Methods for Handling Missing Non-Normal Data in Structural Equation Modeling

By

Fan Jia

Submitted to the graduate degree program in the Department of Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Wei Wu, Chairperson

_____

Pascal Deboeck

_____

Amber Watts

_____

William P. Skorupski

_____

Paul E. Johnson

Date Defended: 05/17/16

The Dissertation Committee for Fan Jia

certifies that this is the approved version of the following dissertation:

Methods for Handling Missing Non-Normal Data in Structural Equation Modeling

_____

Wei Wu, Chairperson

Date approved: 05/17/16

Abstract

Methods for dealing with non-normal data have been broadly discussed in the structural equation modeling (SEM) literature. The issue of how to properly handle normal missing data has also received enough attention. However, much less research has been done to deal with the situation where non-normality and missingness coexist. Generally speaking, there are three classes of methods for dealing with missing non-normal data (continuous and ordinal) in SEM: a) robust procedures, b) Bayesian analysis, and c) multiple imputation. None of these methods, except for robust full information maximum likelihood (robust FIML), have been systematically evaluated in the SEM context with incomplete non-normal data. In this dissertation, I investigated and compared the performance of the three classes of methods under a broad range of conditions for the two types of missing non-normal data.

Acknowledgements

I would like to thank my advisor, Dr. Wei Wu, for all your guidance and encouragement. I appreciate all the training opportunities you have provided and all the wisdom you have shared with me. This work would not have been possible without your continuous support. I greatly appreciate the support and feedback from my committee members, Dr. Paul Johnson, Dr. Pascal Deboeck, Dr. William Skorupski and Dr. Amber Watts. I also want to thank Cathy O'Keefe for your kindness and patience that helped me make critical decisions during the process of my pursuing a Ph.D. I feel fortunate to have met all my fellow students in the Quantitative Psychology program. I will always treasure the moments when we discussed questions, collaborated for research projects and supported each other throughout the program. Finally, to my parents and my husband, Yueqi Yan, thank you for your understanding, caring and love.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Structural equation modeling (SEM) has gained its wide popularity in social and behavioral research in the recent decades. As a powerful multivariate data analysis tool, SEM allows researchers to model latent variables and measurement errors simultaneously. The most popular estimation methods for SEM are maximum likelihood (ML) and generalized least squares (GLS). When certain assumptions are met, ML and GLS possess desirable asymptotic properties, such as unbiasedness, consistency and efficiency (Finney & DiStefano, 2006).

One important assumption is normality. Violation of this assumption can cause problems in parameter estimates and model fit evaluation. More problems may arise in the presence of missing data. Methods to deal with non-normal data have been broadly discussed in the SEM literature. The issue of how to handle normal missing data properly has also received plenty of attention. However, much less research has been done to deal with the situation where non-normality and missingness coexist. When data are incomplete and non-normal, the complete non-normal data methods may need to be adjusted, and the normal-theory-based missing data techniques may also be invalid.

Generally speaking, there are three classes of methods that can be used to handle missing non-normal data: a) robust procedures, b) Bayesian SEM (BSEM), and c) multiple imputation (MI). The purpose of this dissertation is to compare the performance of these methods under a broad range of conditions and find the best method(s) that outperform the others across all conditions. In this Chapter, I provide the background information of the missing non-normal data problem in SEM by introducing two commonly used normal-theory-based estimators, ML and GLS, and the effects of non-normality and missing data on the two estimators. The rest of the dissertation is structured as follows. Chapters 2, 3, and 4 introduce each of the three classes of

methods. In Chapters 5 and 6, I describe two simulation studies that evaluated the performance of these methods when handling missing non-normal continuous data and missing ordinal data, respectively. A variety of influencing factors were taken into consideration in the two studies, such as sample size, degree of non-normality, number of categories, missing data mechanism, and missing data proportion. In Chapter 7, I discuss the findings, limitations, and directions for future research.

Normal-Theory-Based SEM

SEM models are typically comprised of observed variables, latent variables and residuals (Lee, 2007). A structural model can be expressed as

$$\boldsymbol{\eta} = \boldsymbol{\Pi}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{1.1}$$

where $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are the vectors of latent endogenous and exogenous variables, respectively. The residuals are represented by the $\boldsymbol{\zeta}$ vector. The matrices $\boldsymbol{\Pi}$ and $\boldsymbol{\Gamma}$ represent the coefficients for the latent endogenous variables and the exogenous variables, respectively. A structural equation model could have both mean and covariance structures. For simplicity, unless otherwise noted, this dissertation illustrates the rationales of different methods using only covariance structures. The rationales of the methods can be generalized to models with both structures. In SEM, the population covariance matrix of the observed variables $\boldsymbol{\Sigma}$ is formulated as a function the unknown parameters $\boldsymbol{\theta}$ (Bollen, 1989), that is

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) \tag{1.2}$$

One way to estimate the parameters is to minimize the discrepancy between the observed sample covariance matrix $\mathbf{S}$ and the model-implied covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. The discrepancy could be

written as $F = F[\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})]$, which is called the fit function. When the discrepancy is zero, it indicates that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ "fits" the data perfectly. The extent to which $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ fits the data can be assessed by a χ^2 test statistic and other fit indices.

In SEM, the most commonly used estimators are maximum likelihood (ML) and generalized least squares (GLS). The two estimators are referred to as the normal-theory-based estimators, because they are based on the assumption that the data are continuous and multivariate normally distributed. The fit function of ML is given by (Browne, 1984)

$$F_{\text{ML}} = \log|\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})| - \log|\mathbf{S}| + \text{tr}[\mathbf{S}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}] - p \tag{1.3}$$

where p is the number of observed variables, and the normal-theory-based GLS fit function has the following form (Bollen, 1989)

$$F_{\text{NGLS}} = \frac{1}{2}\text{tr}[(\mathbf{S} - \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}))\mathbf{S}^{-1})]^2 \tag{1.4}$$

Most of the time, the minimum of these fit functions cannot be obtained directly, thus some iterative procedures become necessary. A review on the iterative procedures used in the major SEM software packages can be found in Lee (2007, Section 3.6).

When the normality assumption and other assumptions, such as independent observations, large sample size, and correctly specified model (Bollen, 1989; Finney & DiStefano, 2006; Savalei & Falk, 2014), are met, the parameter estimates produced by ML and GLS have desirable asymptotic properties, such as unbiasedness (close enough to the true population values), consistency (converge to the true values as the sample size goes large), and efficiency (sampling distribution of the estimates has minimum variance). In addition, the model test statistic, defined as $T = (N-1)F$ or $T = NF$, follows a χ^2 distribution with $p^* - q$ degrees of

freedom, where p^* is the number of the non-duplicated elements in the observed covariance matrix, $p^* = 0.5p(p+1)$, and q is the number of unknown parameters (Finney & DiStefano, 2006; Lee, 2007).

Violation of Normality

As discussed above, ML and GLS are both based on the assumption of multivariate normality, which also implies that data are continuous. This assumption is very likely to be violated in practice. There are two types of non-normal data: 1) non-normal continuous data and 2) categorical data, such as ordinal and nominal data. This dissertation focuses on non-normal continuous data and ordinal (including ordered binary) data, which are most common in social science. In what follows, the characteristics of the two types of non-normal data and the robustness of the normal-theory-based estimators to the two types of data are discussed.

Non-normal continuous data. For continuous data, non-normality is reflected by the moments around the mean of a distribution (Bollen, 1989). Let **x** be a random variable with a population mean μ, then the r^{th} moment about the mean is computed by $\mu_r = \mathrm{E}[(\mathbf{x}-\mu)^r]$, for $r >$ 1. The most well-known moments are the first moment (mean, i.e., μ_1) and the second moment (variance, i.e., μ_2). The typical indices for non-normality are the standardized third moment (skewness, see Equation [1.5]) and the standardized fourth moment (kurtosis, see Equation [1.6]).

$$skewness = \frac{\mu_3}{(\mu_2)^{3/2}} \tag{1.5}$$

$$kurtosis = \frac{\mu_4}{\mu_2^{\,2}} \tag{1.6}$$

The skewness describes the asymmetry of a distribution about its mean, and the kurtosis measures the "peakedness" of a distribution. For a univariate normal distribution, the skewness is 0 and the kurtosis is 3 (or excess kurtosis = kurtosis – 3 = 0). The departures from the two values (i.e. the degree of non-normality) can be tested by specific statistics (see Bollen, 1989, p.421, Table 9.2, for formulas).

In SEM, the multivariate normality is of greater concern. Even when the marginal distribution of each variable is univariately normal, it is possible that the variables are not multivariate normally distributed. To evaluate multivariate normality, Mardia (1970; 1985) developed a measure of multivariate kurtosis and a test statistic for this measure. Mardia's kurtosis and its associated test statistic are available in major software packages such as Mplus, EQS, R and SAS (Yuan, Lambert, & Fouladi, 2004).

Ordinal data. The assumption of normality is always violated when data are not continuous. Ordinal data, such as the data measured using Likert scales, are very common in social and behavioral science. An ordinal variable contains only a few response points, which are ordered, but the distances among the values are not meaningful. According to Bollen (1989), when the observed indicators in SEM models are ordinal, there are at least two important consequences if the ordinal data are treated as normal: 1) the linear measurement model does not hold for ordinal indicators; and 2) the fundamental hypothesis of SEM does not hold ($\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}(\boldsymbol{\theta})$).

One way to think of how the ordinal data occur is that they are discretized from a continuous latent variable (Bollen, 1989; Finney & DiStefano, 2006; Muthén, 2000). Let $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_j)'$ be a vector of observed ordinal variables, and $\mathbf{Z}^* = (\mathbf{z}_1^*, \mathbf{z}_2^*, ..., \mathbf{z}_j^*)'$ be the vector of the corresponding underlying latent variables, which is categorized as follows.

$$\mathbf{z}_j = c \Leftrightarrow a_{(c\text{-}1)j} < \mathbf{z}_j^* < a_{cj} \tag{1.7}$$

where c is an integral value that indicates the category, $c = 1, 2, \ldots, C$. a_{cj} is the c^{th} threshold for the j^{th} variable ($a_{0j} = -\infty$, and $a_{cj} = +\infty$). When the distributions of $\mathbf{Z}^*$ are known, the thresholds can be estimated. This method assumes that the underlying continuous latent variables $\mathbf{Z}^*$ are normally distributed, although this assumption is very difficult to test (Bentler & Chou, 1987; Muthén, 2000).

Robustness of the normal-theory-based estimators. When data have a non-normal distribution, it is natural to ask whether it is still appropriate to use the normal-theory-based estimators. The robustness of ML and GLS to non-normality and discontinuity has been studied for decades. These studies mostly focus on the influence of non-normality on the parameter estimates, standard errors, χ^2 test statistic and other fit indices.

Browne (1984) provides theoretical proofs for the effects of (continuous) non-normality on these quantities. He found that the consistency of the two estimators holds true even when data are not normal; however, "the test statistics and the estimator standard errors … are inappropriate for any multivariate distribution whose kurtosis differs from that of the normal distribution" (Browne, 1984, p. 63). Empirical studies have supported his findings (Chou, Bentler, & Satorra, 1991; Curran, West, & Finch, 1996; Fan & Wang, 1998; Finch, West, & MacKinnon, 1997; Olsson, Foss, Troye, & Howell, 2000).

The effect of discontinuity on the performance of ML and GLS is dependent on at least two factors: 1) the distribution of categorical variables and 2) the number of categories. Researchers generally hold the opinion that when the ordinal data are not severely skewed or kurtotic, and have at least five categories, treating them as continuous does not result in severe

bias in parameter estimates, standard errors, or fit indices (Finney & DiStefano, 2006). In other situations, in which there is a small number of categories or/and higher levels of skewness and kurtosis, bias in parameter estimates and standard errors could be more pronounced, and the fit indices could be misleading (Dolan, 1994; Green, Akey, Fleming, Hershberger, & Marquis, 1997; Muthén & Kaplan, 1985).

Missing Data

Missing data mechanisms. Missing data are pervasive in social and behavioral science. Conventional methods for dealing with missing data, such as listwise deletion, pairwise deletion, and mean or regression imputation, could cause bias in parameter and standard error estimates and loss in power (Allison, 2000; Enders, 2010; Little & Rubin, 2002). To appropriately handle missing data, one needs to better understand the processes by which data become missing. Rubin (1976) created a classification scheme that describes the processes. If the probability of having missing data on **Y** is not related to the values of **Y** itself, after controlling for the other variables in the analysis, then the data on **Y** are said to be missing at random (MAR). Otherwise, the data are said to be missing not at random (MNAR). A special case of MAR is missing completely at random (MCAR), that is, the probability of missing data on **Y** is unrelated to the values of **Y** itself or any other variables in the data set. MCAR, MAR and MNAR are called the "three missing data mechanisms", which are widely used in the literature. Evidences show that MAR, including MCAR, could be well handled by modern missing data techniques (Enders, 2001a, 2010; Graham, 2009; Rubin, 1976, 1996; Schafer & Graham, 2002), such as full information maximum likelihood and multiple imputation (FIML and MI; both will be introduced in the following sections). Special treatments for MNAR are also available. However, they are more

complex than the MAR techniques and should be used with caution as they do not always perform better than the MAR techniques when handling MNAR data (Enders, 2011).

Missingness and non-normality. FIML and MI assume normality. However, when data are incomplete, their distribution properties, such as skewness and kurtosis, cannot be easily measured. Unless data are MCAR, the observed data in an incomplete data set can possess significant skewness and kurtosis even when the population distribution is normal. Conversely, with MAR or MNAR, the observed data could pass a skewness or kurtosis test when the population distribution is in fact non-normal (Yuan et al., 2004; Yuan, Yang-Wallentin, & Bentler, 2012) .Yuan and colleagues (2004) proposed a procedure that extends the Mardia's kurtosis (1970) to missing data. They noted that the missing data version of the Mardia's kurtosis is "as good as its complete data counterpart" (p. 432) for MCAR or MAR, but caution is needed because it only applies to certain conditions. Likewise, when there are missing values on ordinal variables, the number of categories could be reduced. As a result, the underlying distribution of the observed data could be very different than that in the population. To my knowledge, no index or test has been proposed to evaluate the underlying normality of incomplete ordinal data. Due to the complexity introduced by the coexistence of missingness and non-normality, applied users sometimes may need to rely on the robustness of FIML and MI. Unfortunately, research has shown that FIML and MI do not always work well when treating non-normal data as normal (e.g., Enders, 2001b; Yuan et al., 2012). Thus, there is a strong demand for strategies that can handle both missingness and non-normality appropriately.

Methods for Missing Non-Normal Data: Scope and Significance of the Dissertation

Generally speaking, three classes of methods can be used to handle both missingness and non-normality: a) robust procedures, b) Bayesian SEM (BSEM), and c) multiple imputation (MI).

Among these methods, robust procedures have received the most attention in the SEM literature. Specifically, rescaled ML (robust ML) and diagonally weighted least squares (cat-DWLS) are the most commonly used robust estimators for non-normal continuous data and ordinal data, respectively. In the presence of missing data, the extension of robust ML, i.e., robust FIML, has been found effective with continuous non-normal data (e.g., Enders, 2001b) and has been widely used in practice even when data are ordinal. For ordinal data, cat-DWLS was found to perform well under MCAR, but could cause problems under MAR (Asparouhov & Muthén, 2010a, 2010e).

Bayesian analysis has a long history in statistics, but has just gained popularity in SEM since the start of 20th century. BSEM can appropriately accommodate ordinal data under both MCAR and MAR, although its performance has not been thoroughly investigated. When data are continuous, BSEM generally assumes normality. It is not clear whether BSEM could still produce acceptable results when data are skewed or kurtotic.

The standard MI method assumes multivariate normality. This method is therefore called multivariate normal imputation (MI-MVN). Although MI-MVN was found robust to mild non-normal data (Demirtas et al., 2008), it might yield biased results when data are severely non-normal (Yuan et al., 2012). Regarding ordinal data, Asparouhov & Muthén (2010a, 2010c) found that latent variable imputation (MI-LV) followed by cat-DWLS was superior to the direct cat-DWLS for MAR data in a limited set of conditions. The former strategy needs to be evaluated under a wider range of conditions. Multiple imputation by chained equations (MICE) is also a promising imputation algorithm. It is very flexible in using a variety of imputation models or techniques to deal with various data types and different relations in the data (van Buuren, 2012; van Buuren et al., 2006). MICE imputes data on a variable-by-variable basis, and can work with

both parametric models (e.g., linear regression and logistic regression) and nonparametric techniques (e.g., predictive mean matching and random forests). However, little is known about the performance of the MICE methods in the SEM context.

This dissertation aims to fill the gaps in the literature discussed above. First, I investigated the robustness of the normal-theory-based methods (e.g., MI-MVN and BSEM) in the context of missing continuous non-normal data. For missing ordinal data, I also examined the robustness of the continuous-data methods (e.g., robust FIML and MI-MVN). Second, the missing ordinal data methods that received limited attention in SEM (e.g., cat-DWLS and MI-LV) were thoroughly evaluated in a broader range of scenarios. In addition, the methods that have not been evaluated in the context of SEM (e.g., the MICE methods) were explored.

Chapter 2: Robust Procedures

The most well-known strategy to deal with non-normality is to correct for the inappropriate standard errors and the test statistic. Methods referred to as robust corrections have been developed based on this idea (Savalei, 2014). These methods have also been extended to the situations in which missing data are present. This chapter explains the basic idea of the robust procedures (also known as rescaling procedures) for complete continuous data, followed by the extension to non-normal continuous data and ordinal data. The performance of the robust procedures in the literature is then reviewed.

The General Idea

General forms of standard errors and the test statistic. As discussed above, the basic hypothesis of SEM is that the population covariance matrix of the observed variables is a function of the unknown parameters in a vector $\boldsymbol{\theta}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ (Bollen, 1989). Let **S** and **R** represent the $p \times p$ sample covariance matrix and the $p \times p$ residual variance matrix, respectively, then

$$\mathbf{S} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) + \mathbf{R} \tag{2.1}$$

The matrices in Equation (2.1) can be also vectorized to define a model equation analogous to regular regression. Let vec($\cdot$) be the vectorizing function that turns a $p \times p$ matrix to a $p^* \times 1$ vector by stacking the lower triangle elements of the matrix, row by row sequentially, $p^* = 0.5p(p+1)$, then Equation (2.1) can be rewritten as (Browne, 1984; Savalei, 2014):

$$\mathbf{s} = \boldsymbol{\sigma}(\boldsymbol{\theta}) + \mathbf{r} \tag{2.2}$$

where **s** = vec(**S**), **r** = vec(**R**), and $\boldsymbol{\sigma}(\boldsymbol{\theta})$ is a nonlinear function of the model parameters $\boldsymbol{\theta}$. Based on Equation (2.2) , the SEM fit function can be written as (Savalei, 2014)

$$F = (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}))' \mathbf{W}^{-1} (\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta})) \tag{2.3}$$

where $\mathbf{W}$ is a positive defined $p^* \times p^*$ weight matrix. To obtain an efficient F , $\mathbf{W}$ has to be a consistent estimator of the asymptotic covariance matrix of the elements in **s**.

The estimates of $\boldsymbol{\theta}$ is obtained by minimizing the fit function (2.3) iteratively using optimization algorithms. The asymptotic covariance matrix of the parameter estimates, also known as the inverse of Fisher information matrix, is given by

$$\text{cov}(\sqrt{N}\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\Delta}}' \mathbf{W}^{-1} \hat{\boldsymbol{\Delta}})^{-1} \tag{2.4}$$

where $\hat{\boldsymbol{\Delta}} = \frac{\partial \boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big|_{\hat{\boldsymbol{\theta}}}$ is the $p^* \times q$ matrix of the first derivatives of $\boldsymbol{\sigma}(\boldsymbol{\theta})$ evaluated at the parameter estimates $\hat{\boldsymbol{\theta}}$ (Savalei, 2014).

Weight matrices in normal-theory-based estimators. Under the assumption of multivariate normality, the asymptotic covariance matrix of **s** has a simple form, denoted as Γ. A typical element of Γ is given by (Browne, 1984; Savalei, 2014)

$$\Gamma_{ij,kl} = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} \tag{2.5}$$

where σ_{ik} , σ_{jl} , σ_{il} , and σ_{jk} are elements of the population covariance matrix, which can be easily estimated from the raw data. For example, the normal-theory-based ML and GLS use the model implied covariance matrix and the sample covariance matrix, respectively, to estimate the population covariance matrix, and therefore they are asymptotically equivalent (Bollen, 1989). Substituting $\mathbf{W}$ by Γ in Equation (2.4) and (2.3), the appropriate standard errors and the

asymptotic χ^2 distributed test statistic ($T = N\hat{F}$) can be computed. However, when the normality assumption is violated, using Γ in the fit function would produce incorrect standard errors and a misleading test statistic. To correct for non-normality, a better estimate for the asymptotic covariance matrix should be obtained. Various methods have been developed for this purpose. These methods could also be extended to the missing data context.

Weighted least squares. To correct for non-normality, one strategy is to use an estimator which reflects the true asymptotic covariance matrix of **s** in non-normal data. Browne (1984) developed an asymptotically distribution-free estimator for continuous data (ADF, also known as weighted least squares or WLS), which, as its name indicates, does not require the assumption of multivariate normality. In ADF, the asymptotic covariance matrix $\mathbf{W}$ in Equation (2.4) and (2.3) is defined as $\mathbf{\Gamma}^*$, the typical element of which is given by

$$\Gamma^*_{ij,kl} = \sigma_{ijkl} - \sigma_{ij}\sigma_{kl} \tag{2.6}$$

where σ_{ijkl} is the fourth moment of the data. The ADF estimate of $\Gamma^*_{ij,kl}$ is $\hat{\Gamma}^*_{ij,kl} = s_{ijkl} - s_{ij}s_{kl}$, which can be obtained from the observed data. ADF is statistically efficient but can be computationally intensive when $\mathbf{\Gamma}^*$ is large (Finney & DiStefano, 2006), as $\hat{\mathbf{\Gamma}}^*$ needs to be inversed when minimizing the fit function and computing standard errors (see Equation [2.4] and [2.3]). In addition, it requires a very large sample size to produce stable estimates. Hu, Bentler, and Kano (1992) and Curran et al. (1996) found that only with extremely large sample size (e.g., N = 5000), ADF could produce a reliable χ^2 test statistic. With medium and small sample sizes, ADF produced biased parameter estimates and standard errors, as Hoogland and Boomsma (1998) point out.

When data are ordinal, a strategy is to fit the SEM model to the polychoric correlation matrix rather than the sample covariance matrix. Polychoric correlations represent the relations between the continuous latent variables underlying the ordinal data. Therefore, in the SEM fit function, **s** represents the vector that contains $p^{**} = 0.5p(p-1)$ unduplicated elements of the sample polychoric correlations, and $\boldsymbol{\sigma(\theta)}$ is the vector of the residual polychoric correlations , which has the same dimension of **s** (see Equation [2.3]). When the weight matrix $\mathbf{W}$ in the fit function is defined as the ADF type asymptotic covariance matrix of polychoric correlations ($\boldsymbol{\Gamma}_c^*$), we have a categorical data version of WLS, called cat-WLS. Similar to ADF, this method is efficient and produces asymptotically χ^2 distributed test statistic; however, it requires a very large sample size, and therefore it is not usually recommended (DiStefano, 2002; Dolan, 1994; Hoogland & Boomsma, 1998).

Solutions to Non-Normal Continuous Data: Rescaling

Rescaled ML (robust ML). One strategy has been developed to avoid the intensive computation of the inversing of the asymptotic covariance matrix. The main idea is to use a simpler estimator accompanied with robust corrections. Satorra and Bentler (1994) developed a method that is based on the normal-theory estimates but corrects for the impact of non-normality by rescaling the standard errors and the test statistic. This method is known as Satorra-Bentler scaling.

The Satorra-Bentler scaling is most commonly applied to ML, although it theoretically can be applied to any of the normal-theory estimators. Assuming multivariate normality, the ML estimator minimizes the following discrepancy function

$$F_{\mathrm{ML}} = \log|\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})| - \log|\mathbf{S}| + \mathrm{tr}[\mathbf{S}\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}] - p \tag{2.7}$$

Shapiro (1985) shows that the Equation (2.7) is asymptotically equivalent to Equation (2.3) with $\mathbf{W} = \boldsymbol{\Gamma}$ (also see Equation [2.5]). The chi-square test statistic of ML is then defined as

$$T_{\mathrm{ML}} = N\hat{F}_{\mathrm{ML}} \tag{2.8}$$

According to Satorra and Bentler (1994), the robust chi-square test statistic is given by

$$T_{\mathrm{SB}} = \frac{p^{*} - q}{\mathrm{tr}(\hat{\mathbf{U}}\hat{\boldsymbol{\Gamma}}^{*})} T_{\mathrm{ML}} \tag{2.9}$$

In Equation (2.9), $\hat{\mathbf{U}} = \mathbf{W}^{-1} - \mathbf{W}^{-1}\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}})^{-1}\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}$, where $\mathbf{W} = \hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Delta}} = \frac{\partial \boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big|_{\hat{\boldsymbol{\theta}}}$. This method adjusts T_{ML}, so that the mean of T_{SB} is equal to the mean of a $\chi^2_{p^*-q}$. The distribution of T_{SB} is not exactly χ^2, but it works well, approximately (Savalei, 2014).

The standard errors are also need to be rescaled. Recall that the under the multivariate normality assumption, the covariance matrix of parameters is given by Equation (2.4), while under non-normality, the robust covariance matrix of parameters has a sandwich-like form, as shown in Equation (2.10),

$$\mathrm{cov}(\sqrt{N}\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}})^{-1}\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Gamma}}^{*}\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}})^{-1} \tag{2.10}$$

where $\mathbf{W} = \hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Delta}} = \frac{\partial \boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big|_{\hat{\boldsymbol{\theta}}}$. In Equation (2.10), the inverse of Fisher information matrix $(\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}})^{-1}$, which is equivalent to Equation (2.4), forms the outside "bread" of the sandwich, and middle part(i.e. the "meat") is $\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Gamma}}^{*}\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}}$, where $\hat{\boldsymbol{\Gamma}}^{*}$ is the same as the asymptotic covariance matrix of ADF and can be computed based on Equation (2.6). Thus, this covariance

matrix is also called the "sandwich covariance matrix". When the multivariate normality holds, $\hat{\boldsymbol{\Gamma}}^{*} = \hat{\boldsymbol{\Gamma}}$ and Equation (2.10) is reduced to $(\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}})^{-1}$, otherwise the "meat" part $\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Gamma}}^{*}\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}}$ serves as a correction factor that influences the magnitudes of $\text{cov}(\sqrt{N}\hat{\boldsymbol{\theta}})$, depending on the kurtosis of the data (Enders, 2010). For example, in a leptokurtic distribution (i.e., excess kurtosis > 0), the existence of the extreme scores increases the values in $\hat{\boldsymbol{\Gamma}}^{*}$ relative to $\hat{\boldsymbol{\Gamma}}$, the corresponding robust standard errors are greater than the normal-theory- based standard errors. On the contrary, when the distribution is platykurtic (i.e., excess kurtosis < 0), the robust standard errors become smaller than the normal-theory-based standard errors.

In Equation (2.10), $\hat{\boldsymbol{\Gamma}}^{*}$ does not need to be inversed, so that it saves the computational efforts drastically, especially when the model is complex.

Rescaled FIML (robust FIML). The idea of a sandwich covariance matrix is also applicable when data are incomplete. The difference is that the sandwich covariance matrix is now estimated from a missing data estimator.

When data are incomplete but normally distributed, SEM estimates can be obtained by iteratively maximizing the sum of *N* casewise log-likelihood functions (Enders, 2001a),

$$\log l_{\text{FIML}} = K - \frac{1}{2}\sum_{i}^{N}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i}^{N}(\mathbf{X}_{i,\text{obs}} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{X}_{i,\text{obs}} - \boldsymbol{\mu}_i) \tag{2.11}$$

where K is a constant. The subscript i associated with the covariance matrix ($\boldsymbol{\Sigma}$) and the mean vector ($\boldsymbol{\mu}$) indicates the fact that the dimension of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ vary across cases. This method is referred to as full information maximum likelihood (FIML). FIML is the most commonly used missing data technique. Under the assumption of MAR and normality, FIML is known to produce unbiased and efficient parameter estimates and correct χ^2 test statistic (Enders &

Bandalos, 2001). However, similar to ML estimator with complete data, when data are non-normal, FIML standard errors tend to be negatively biased and the χ^2 test statistic is overestimated (Enders, 2001b).

To deal with missing non-normal data, Yuan and Bentler (2000) have extended the Satorra-Bentler method to correct the FIML standard errors and the test statistic under the assumption of MCAR. The asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by

$$\text{cov}(\sqrt{N}\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\Omega}} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1} \tag{2.12}$$

In Equation (2.12), $\mathbf{A} = -\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{\partial^2 \mathbf{l}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$, known as the observed information matrix, and $\mathbf{B} = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \mathbf{l}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \frac{\partial \mathbf{l}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}$. The derivatives in both **A** and **B** are evaluated at $\hat{\boldsymbol{\theta}}$, and $\mathbf{l}_i(\boldsymbol{\theta})$ is the log-likelihood for case i for the structured model. Similar to its counterpart in the complete data context, in this "sandwich covariance matrix", $\hat{\mathbf{A}}^{-1}$ forms the "bread" part, and $\hat{\mathbf{B}}$ is the "meat". When data are normal, $\hat{\mathbf{B}} = \hat{\mathbf{A}}$, which reduces the asymptotic covariance of $\hat{\boldsymbol{\theta}}$ to the inverse of observed information matrix $\hat{\mathbf{A}}^{-1}$. With non-normal data, $\hat{\mathbf{B}}$ reflects the kurtosis of the observed data, and corrects the standard errors.

To obtain the robust FIML test statistic, first let $\boldsymbol{\beta}$ represent the vector of model parameters under the saturated (unstructured) model. The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is defined as (Savalei & Bentler, 2005; Yuan & Bentler, 2000)

$$\text{cov}(\sqrt{N}\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\Omega}}_{\beta} = \hat{\mathbf{A}}_{\beta}^{-1}\hat{\mathbf{B}}_{\beta}\hat{\mathbf{A}}_{\beta}^{-1} \tag{2.13}$$

In Equation (2.13), $\mathbf{A}_{\beta} = -\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \mathbf{l}_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}$ and $\mathbf{B}_{\beta} = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \mathbf{l}_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\frac{\partial \mathbf{l}_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'}$, both are evaluated at $\boldsymbol{\beta}(\hat{\boldsymbol{\theta}})$, and $\mathbf{l}_i(\boldsymbol{\beta})$ is log-likelihood for case i for the saturated model. Then the FIML test statistic can be rescaled as

$$T_{\text{RFIML}} = \frac{p^* - q}{\text{tr}(\hat{\boldsymbol{\Omega}}_{\beta}\hat{\mathbf{U}})} T_{\text{FIML}} \tag{2.14}$$

In Equation (2.14), $\hat{\mathbf{U}} = \hat{\mathbf{A}}_{\beta} - \hat{\mathbf{A}}_{\beta}\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\Delta}}'\hat{\mathbf{A}}_{\beta}\hat{\boldsymbol{\Delta}})^{-1}\hat{\boldsymbol{\Delta}}'\hat{\mathbf{A}}_{\beta}$, where $\hat{\boldsymbol{\Delta}} = \frac{\partial\boldsymbol{\beta}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\Big|_{\hat{\boldsymbol{\theta}}}$, the matrix of model derivatives evaluated at the FIML estimate $\hat{\boldsymbol{\theta}}$; $T_{\text{FIML}} = -2[\mathbf{l}(\hat{\boldsymbol{\theta}}) - \mathbf{l}(\hat{\boldsymbol{\beta}})]$, where $\mathbf{l}(\hat{\boldsymbol{\theta}})$ is the maximized log-likelihood under the structured model, and $\mathbf{l}(\hat{\boldsymbol{\beta}})$ is the corresponding log-likelihood of the saturated model. As the complete-data robust test statistic, the asymptotic distribution of T_{RFIML} approximates only the mean of χ^2 square distribution, but it can be used as an approximation of χ^2 (Savalei & Bentler, 2005; Savalei & Falk, 2014).

Solutions to Ordinal Data: Simplified WLS-Type Estimators

Diagonally weighted and unweighted least squares (cat-DWLS and cat-ULS). As discussed above, when data are ordinal, one strategy is to fit the SEM model to the polychoric correlation matrix rather than the sample covariance matrix, using a weighted-least-square-type (WLS-type) estimator. Three major methods have been developed. One is called cat-WLS (i.e., WLS for categorical data). Similar to its counterpart in the complete data context (i.e., ADF), this method requires vary large sample size (DiStefano, 2002; Dolan, 1994; Hoogland & Boomsma, 1998). To solve this problem, researchers have proposed to use simpler estimators and adjust the standard errors and the test statistic using robust corrections. One of the estimators is called

diagonally weighted least squares (cat-DWLS), in which the weight matrix in the fit function is the diagonal of $\boldsymbol{\Gamma}_c^*$, i.e., $\boldsymbol{\Gamma}_{diagc}^* = \text{diag}(\boldsymbol{\Gamma}_c^*)$. Another estimator is called unweighted least squares (cat-ULS) estimator, which has an identity "weight" matrix. The robust standard errors can be computed based on Equation (2.10) by substituting $\hat{\boldsymbol{\Gamma}}^*$ by $\hat{\boldsymbol{\Gamma}}_c^*$, and use $\hat{\boldsymbol{\Gamma}}_{diagc}^*$ or $\mathbf{I}$ for $\mathbf{W}$. The following equations show the covariance matrices of the parameters for cat-DWLS and cat-ULS, respectively:

$$\text{cov}(\sqrt{N}\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}})^{-1}\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Gamma}}_c^*\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}}'(\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}})^{-1} \tag{2.15}$$

where $\mathbf{W} = \hat{\boldsymbol{\Gamma}}_{diagc}^*$, and

$$\text{cov}(\sqrt{N}\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\Delta}}'\hat{\boldsymbol{\Delta}})^{-1}\hat{\boldsymbol{\Delta}}'\boldsymbol{\Gamma}_c^*\hat{\boldsymbol{\Delta}}'(\hat{\boldsymbol{\Delta}}'\hat{\boldsymbol{\Delta}})^{-1} \tag{2.16}$$

Muthén, du Toit, and Spisic (1997) developed two ways to compute the robust χ^2 test statistic for cat-DWLS: 1) mean-adjusted χ^2 and 2) mean- and variance-adjusted χ^2. A mean-adjusted χ^2 is defined as

$$T_{\text{DWLS-M}} = \frac{d}{\text{tr}(\hat{\mathbf{U}}\hat{\boldsymbol{\Gamma}}_c^*)}(N\hat{F}_{\text{cat-DWLS}}) \tag{2.17}$$

where $d = p^{**} - q$, and $\hat{\mathbf{U}} = \mathbf{W}^{-1} - \mathbf{W}^{-1}\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}\hat{\boldsymbol{\Delta}})^{-1}\hat{\boldsymbol{\Delta}}'\mathbf{W}^{-1}$, where $\mathbf{W} = \hat{\boldsymbol{\Gamma}}_{diagc}^*$. A mean- and variance-adjusted χ^2 (denoted as $T_{\text{DWLS-MV}}$) has the same form as Equation (2.17), the only difference is that in $T_{\text{DWLS-MV}}$, d is computed as the integer closest to d^*,

$$d^* = \frac{[\text{tr}(\hat{\mathbf{U}}\hat{\boldsymbol{\Gamma}}_c^*)]^2}{\text{tr}[(\hat{\mathbf{U}}\hat{\boldsymbol{\Gamma}}_c^*)^2]} \tag{2.18}$$

Asparouhov and Bengt Muthén (2010d) proposed a new way to compute the mean- and variance-adjusted χ^2 (denoted as $T_{\text{DWLS-MV2}}$). They noted that $T_{\text{DWLS-MV2}}$ has a theoretical advantage over $T_{\text{DWLS-MV}}$, because it uses the usual degrees of freedom d rather the estimated d^*, which simplifies the computation of the difference in degrees of freedom when comparing two nested models.

$$T_{\text{DWLS-MV2}} = aT_{\text{DWLS}} - b \tag{2.19}$$

where $a = \sqrt{\frac{d}{\text{tr}[(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}}_{\text{c}}^{*})^2]}}$, and $b = d - \sqrt{\frac{d[\text{tr}(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}}_{\text{c}}^{*})]^2}{\text{tr}[(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}}_{\text{c}}^{*})^2]}}$. The cat-ULS version of the test statistic can be formed in a similar way.

Cat-DWLS and cat-ULS with missing data. Asparouhov and Muthén (2010e) described how these estimators can be modified to accommodate missing data. Basically, these methods use pairwise deletion to estimate polychoric correlations. The resulting polychoric correlations are then used in the SEM fit function, as discussed above. These methods only work well under MCAR and a special case of MAR (i.e., only the covariates have an effect on the missing data patterns). Under a more general assumption of MAR (i.e., both the covariates and the observed dependent variables in each pattern have an effect on the missing data patterns), the incomplete variable (**y**) and the variable that relates to the missingness on **y** should be simultaneously estimated; otherwise the estimates will be biased.

Performance of the Robust Procedures in the Literature

In terms of estimators for continuous data, Enders (2001b) notes that robust FIML "may nearly eliminate the negative impact of non-normal missing data" under both MCAR and MAR. However, Savalei and Falk (2014) found that under a certain type of MAR (the type which

mainly occurred on the heavy tail of distribution), robust FIML could perform poorly when the proportion of missing data was large (30%). In addition, although robust FIML was developed for missing continuous data, it has been widely used to deal with missing categorical data in practice. However, the literature on the performance of robust FIML for categorical incomplete data is limited.

Turning to estimators for categorical data, Brown (2006) suggests that cat-DWLS is the "best" estimator for complete ordinal data. Forero, Maydeu-Olivares, and Gallardo-Pujol (2009) show that cat-DWLS and cat-ULS produced very similar results, and the robust standard errors associated with cat-ULS slightly outperformed those of cat-DWLS. However, the missing data versions of cat-DWLS and cat-ULS were considered unreliable when data were MAR. Asparouhov and Muthén (2010a) evaluated the performance of cat-DWLS in analyzing binary data in three models: a bivariate probit model, a two-level model, and a growth curve model. They found that under MAR, cat-DWLS produced biased parameter estimates in all of the examined conditions. They also noted that this problem may be solved by the Bayes estimator or MI followed by the WLS-type estimators (Asparouhov & Bengt Muthén, 2010a, 2010c).

Chapter 3: Bayesian SEM

Although the robust procedures alleviate the problems in standard error estimates and the test statistic, they may create convergence problems when models are complex and data are incomplete (Lee, 2007). Moreover, when data are categorical and MAR, the robust procedures tend to provide biased parameter estimates. A possible solution to this problem is Bayesian analysis.

Bayesian analysis has been available for decades. However, it only received attention from the SEM users starting in the 20th century (Lee, 2007). Due to the rapid development in computational algorithms based on the Markov chain Monte Carlo (MCMC) approach and the increasing demand of complex models, the Bayesian approaches have been expanded to the SEM framework and are becoming increasingly popular (Kaplan & Depaoli, 2012). According to Lee (2007), estimating structural models through the Bayesian perspective (BSEM) has numerous advantages. First, Bayesian analysis is capable of producing better results than frequentist approaches by allowing the incorporation of true prior information to the observed data. Second, Bayesian models converge more easily due to direct use of raw individual data rather than summary statistics. Third, Bayesian methods depend less on asymptotic theory, and hence provide more reliable results even with non-normal or discrete data. Lastly, missing data can be easily handled using the Bayesian-based algorithms, such as data augmentation (DA; Tanner & Wong, 1987) and the Gibbs sampler (Geman & Geman, 1984). This chapter first introduces the general idea of BSEM, and then discusses how BSEM handles non-normal and missing data, and its performance in the literature.

The General Idea

Bayes' rule. The foundation of Bayesian analysis is the well-known Bayes' rule. Let $\boldsymbol{\theta}$ be the unknown parameters in a model and $\mathbf{Y}$ represent the observed data. The Bayes' rule is given by

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{Y})} \tag{3.1}$$

where $p(\boldsymbol{\theta} \mid \mathbf{Y})$ is the probability of $\boldsymbol{\theta}$ given data $\mathbf{Y}$, which is referred to as the posterior distribution of $\boldsymbol{\theta}$; $p(\boldsymbol{\theta})$ is the prior distribution of the unknown parameters $\boldsymbol{\theta}$; $p(\mathbf{Y} \mid \boldsymbol{\theta})$ is the probability of observing $\mathbf{Y}$ given a set of parameter values $\boldsymbol{\theta}$; and $p(\mathbf{Y} \mid \boldsymbol{\theta})$ is equivalent to the likelihood of $\boldsymbol{\theta}$ given fixed values of $\mathbf{Y}$, which is denoted as $l(\boldsymbol{\theta} \mid \mathbf{Y})$. The marginal probability $p(\mathbf{Y})$ in the denominator is a scaling constant which is used to set the total area of the posterior distribution to be one. Because $p(\mathbf{Y})$ does not involve any model parameters, we can ignore it and use "$\propto$" (interpreted as "is proportional to") to replace the equal sign, which leads to

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) = l(\boldsymbol{\theta} \mid \mathbf{Y})\, p(\boldsymbol{\theta}) \tag{3.2}$$

Priors. Bayesian analysis is under the assumption that the model parameters $\boldsymbol{\theta}$ are random and each possesses a probability distribution, called the prior distribution of $\boldsymbol{\theta}$. This is the key difference between the Bayesian analysis and the standard frequency-and-hypothesis-testing-based analysis (known as frequentist analysis), because the latter assumes that $\boldsymbol{\theta}$ is unknown but fixed (Kaplan & Depaoli, 2012).

Depending on how much information on the distribution of $\boldsymbol{\theta}$ is known before data collection, we can choose one of the two types of priors: 1) non-informative priors or 2) informative priors. When there is little prior knowledge about the distribution, non-informative

priors are often used to "quantify our ignorance" (Kaplan & Depaoli, 2012, p. 652). A non-informative prior could be a uniform distribution with a reasonable range of values or some other distribution with an extremely large variance. As a result, the non-informative priors tend to have less influence on the shape of the posterior distribution, and the Bayesian estimation is mainly affected by the observed data.

However, in many situations, we may have some information about the shape and the scale of $\boldsymbol{\theta}$. This prior information may come from subjective knowledge of field experts or results of closely related studies (Lee, 2007). The knowledge could be incorporated into model estimation through informative priors. Most of the time, an informative prior distribution has its own parameters, which are called hyperparameters.

One important type of informative priors is called "conjugate prior" distributions. A conjugate prior, when combined with the likelihood, produces a posterior distribution that belongs to the same distribution family as the prior itself. For instance, if an observed variable **y** follows a normal distribution, $N(\pi, \sigma^2)$, a conjugate prior distribution of π given a known σ^2 should have an exponential of a quadratic form of π, e.g., $p(\pi) \propto \exp[-\frac{1}{2\tau_0^2}(\pi - \mu_0)^2]$, where μ_0 and τ_0^2 are the hyperparameters of the distribution of π (see Lee, 2007, pp. 72-76, for more details). The benefit of using conjugate priors is that the resulting posterior distribution would have a simple form to solve analytically (Kaplan & Depaoli, 2012).

Posterior analysis. After choosing priors and having the data observed, the posterior distribution could be computed. The Bayesian estimate of $\boldsymbol{\theta}$ is usually obtained by taking the mode or the mean of the posterior distribution, $p(\boldsymbol{\theta} \mid \mathbf{Y})$. The mode (i.e., maximum) of $p(\boldsymbol{\theta} \mid \mathbf{Y})$ could be achieved by using an iterative procedure along with the EM algorithm (Dempster, Laird,

& Rubin, 1977). The mean of the posterior distribution, which is more commonly used, is obtained by the Markov chain Monte Carlo (MCMC) methods.

Markov chain Monte Carlo (MCMC) refers to a class of sampling algorithms to estimate expectations of statistics in a complex model based on the simulation of Markov chains (Gilks, 2005). A Markov chain is a sequence of random elements, $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots,$ for which the distribution of the current element $\boldsymbol{\theta}^{(t)}$ depends on all the previous $\boldsymbol{\theta}$s only through its immediate processor $\boldsymbol{\theta}^{(t-1)}$ (Geyer, 2011; SAS Institute Inc., 2010). For the purpose of demonstration, this dissertation focuses on a MCMC algorithm called the Gibbs sampler (Geman & Geman, 1984). Let q be the number of parameters in $\boldsymbol{\theta}$, then choose a set of starting values for the parameters, that is $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_q^{(0)})$. Based on the observed data $\mathbf{Y}$ and the starting point $\boldsymbol{\theta}^{(0)}$ the Gibbs sampler creates a sampling distribution of $\boldsymbol{\theta}$ by iteratively generating $\boldsymbol{\theta}^{(t)}$ from the conditional distribution $p(\boldsymbol{\theta} \mid \mathbf{Y})$ (Lee, 2007).

$$\begin{aligned} \theta_1^{(t)} &\text{ from } p(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, \ldots, \theta_q^{(t-1)}, \mathbf{Y}) \\ \theta_2^{(t)} &\text{ from } p(\theta_2 \mid \theta_1^{(t-1)}, \theta_3^{(t-1)}, \ldots, \theta_q^{(t-1)}, \mathbf{Y}) \\ &\vdots \\ \theta_q^{(t)} &\text{ from } p(\theta_q \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, \ldots, \theta_{q-1}^{(t-1)}, \mathbf{Y}) \end{aligned} \tag{3.3}$$

where $t = 1, 2, \ldots, T$ denotes the Monte Carlo iterations. Under some general conditions, the sampling distribution of $\boldsymbol{\theta}$ converges at a target distribution as T increases (Geman & Geman, 1984; Gilks, 2005). An initial portion of the Monte Carlo iterations needs to be discarded (termed burn-in iterations). Because the early draws before convergence would be far away from the target distribution. Usually, the number of iterations after burn-in should be a large number, so that the Markov chain can achieve convergence. Multiple chains sampling from different sets of starting points could help achieve convergence with a comparatively small number of

iterations. After MCMC converges, the mean, mode, variance and quartiles of $\boldsymbol{\theta}$ could be obtained based on the posterior distribution (Geyer, 1992; Lee, 2007).

The above demonstration assumes complete data. Missing data, however, can occur in many situations: 1) in factor analysis and SEM, the latent constructs are not directly observed; 2) when observed data are binary or ordinal, usually we assume there are unobserved continuous measurements underlying the categorical data; 3) missing data due to various reasons are very common in practice (Lee, 2007). In Bayesian analysis, the idea of data augmentation (DA; Tanner & Wong, 1987) with MCMC can be used to deal with unobserved data. Treating all the unobserved quantities as missing data, the predictive probability distribution of the missing data conditional on the model parameters is defined. Random values are then drawn from the predictive distribution, which augment with the observed data to define the posterior distribution of model parameters (Lee, 2007). The following sections mainly use the Gibbs sampler as an example to discuss how the Bayesian methods deal with different types of unobserved quantities.

Bayesian estimation of models with latent variables. Different from standard SEM, which are based on mean vectors and covariance matrices, Bayesian analysis requires raw data. When a model contains latent variables, the observed data are augmented with the latent variables, which are treated as hypothetical missing data. Following this data augmentation step, the MCMC algorithms draw samples from the conditional density functions of the unknown parameters and the missing latent values. Taking the Gibbs sampler for example, let $\mathbf{Y}$ be the observed data set with sample size *N*, and $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ represent the unknown parameters and latent variables, respectively. Based on some starting values of $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$, $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_q^{(0)})$ and $\boldsymbol{\Omega}^{(0)} = (\omega_1^{(0)}, \omega_2^{(0)}, ..., \omega_r^{(0)})$, observations of each component in $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ are iteratively drawn from their probability conditional on the values of all the other components and the data $\mathbf{Y}$ (Lee, 2007).

Each iteration includes a sampling procedure for the $\boldsymbol{\theta}$ components (similar to [3.3]) and a sampling procedure for the $\boldsymbol{\Omega}$ components.

$$
\begin{gathered}
\text{Step 1} \\
\theta_1^{(t)} \text{ from } p(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, ..., \theta_q^{(t-1)}, \boldsymbol{\Omega}^{(t-1)}, \mathbf{Y}) \\
\theta_2^{(t)} \text{ from } p(\theta_2 \mid \theta_1^{(t-1)}, \theta_3^{(t-1)}, ..., \theta_q^{(t-1)}, \boldsymbol{\Omega}^{(t-1)}, \mathbf{Y}) \\
\vdots \\
\theta_q^{(t)} \text{ from } p(\theta_q \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, ..., \theta_{q-1}^{(t-1)}, \boldsymbol{\Omega}^{(t-1)}, \mathbf{Y}) \\
\text{Step 2} \\
\omega_1^{(t)} \text{ from } p(\omega_1 \mid \boldsymbol{\theta}^{(t)}, \omega_2^{(t-1)}, \omega_3^{(t-1)}, ..., \omega_r^{(t-1)}, \mathbf{Y}) \\
\omega_2^{(t)} \text{ from } p(\omega_2 \mid \boldsymbol{\theta}^{(t)}, \omega_1^{(t-1)}, \omega_3^{(t-1)}, ..., \omega_r^{(t-1)}, \mathbf{Y}) \\
\vdots \\
\omega_r^{(t)} \text{ from } p(\omega_r \mid \boldsymbol{\theta}^{(t)}, \omega_1^{(t-1)}, \omega_2^{(t-1)}, ..., \omega_{r-1}^{(t-1)}, \mathbf{Y})
\end{gathered}
\tag{3.4}
$$

The number of iterations after burn-in should be a sufficiently large number, and based on the large sample of $\boldsymbol{\theta}$ the Bayesian estimate of $\boldsymbol{\theta}$ (i.e. mean of the sample) could be easily computed. Statistical inference about $\boldsymbol{\theta}$ can be conducted via standard methods after obtaining the standard error of $\hat{\boldsymbol{\theta}}$ (i.e. standard deviation of the sample).

In (3.4), there are two types of conditional distributions, $p(\boldsymbol{\Omega} \mid \boldsymbol{\theta}, \mathbf{Y})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{\Omega}, \mathbf{Y})$. With data augmentation, a latent variable model can be treated as a regression model, and when $\boldsymbol{\theta}$ is known, the conditional distribution of $\boldsymbol{\Omega}$ given $\boldsymbol{\theta}$ and $\mathbf{Y}$ can be directly computed (see Lindley and Smith, 1972). The posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{\Omega}$ and $\mathbf{Y}$ is proportional to $p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\Omega}) p(\boldsymbol{\Omega} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$ according to the Bayes' rule. Thus, the priors for different types of parameters in $\boldsymbol{\theta}$ need to be specified. Lee (2007) provides some possible conjugate priors for SEM models. For example, a residual variance follows an inverse-gamma distribution, the

distribution of the corresponding loading is normally distributed, and the variance and covariance matrix follow an inverse-Wishart distribution (see Lee, 2007, pp. 75-76).

BSEM with Non-Normal Data

The standard BSEM also assumes normality. Despite the rapidly increasing popularity of BSEM methods, little is known about the robustness of the method to non-normal continuous data. Some procedures have been discussed in recent years, such as the multivariate t-distribution method and the P-spline transformation method (Lee & Song, 2004b, 2012; Lee & Xia, 2008). However, these methods are either too dependent on the selection of prior or too difficult to implement, thus, they have limited applications and have not been widely studied.

Using the Bayesian method to deal with ordinal data is less complicated. One popular way is to treat an ordinal variable as the proxy of an underlying normal continuous variable, as discussed above. MCMC is implemented assuming the underlying normal continuous variables are missing data.

The posterior analysis for SEM with ordinal observed data takes three steps. The first two steps are similar to (3.4), that is sampling $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ from their fully conditional distributions. The only difference is that with ordinal data, the underlying latent variables and their thresholds also need to be considered. Let **Y** and **Z** indicate the observed continuous and ordinal variables, respectively, with sample size *N*. Also let $\boldsymbol{\theta}$ given $\boldsymbol{\Omega}$ represent the unknown parameters and the latent variables, respectively. The underlying latent variables and their thresholds are denoted as $\mathbf{Z}^*$ and **A**. Based on the starting values $(\boldsymbol{\theta}^{(0)}, \boldsymbol{\Omega}^{(0)}, \mathbf{Z}^{*(0)}, \mathbf{A}^{(0)})$, observations of each component are iteratively drawn conditionally on the values of all the other components and the observed data. The three steps of this procedure are as follows (Lee, 2007):

$$\begin{gathered}\text{Step 1} \\ \boldsymbol{\theta}^{(t)} \text{ from } p(\boldsymbol{\theta} \mid \boldsymbol{\Omega}^{(t-1)}, \mathbf{Z}^{*(t-1)}, \mathbf{A}^{(t-1)}, \mathbf{Y}, \mathbf{Z}) \\ \text{Step 2} \\ \boldsymbol{\Omega}^{(t)} \text{ from } p(\boldsymbol{\Omega} \mid \boldsymbol{\theta}^{(t)}, \mathbf{Z}^{*(t-1)}, \mathbf{A}^{(t-1)}, \mathbf{Y}, \mathbf{Z}) \\ \text{Step 3} \\ (\mathbf{Z}^{*(t)}, \mathbf{A}^{(t)}) \text{ from } p(\mathbf{Z}^{*}, \mathbf{A} \mid \boldsymbol{\theta}^{(t)}, \boldsymbol{\Omega}^{(t)}, \mathbf{Y}, \mathbf{Z})\end{gathered} \tag{3.5}$$

After achieving convergence, the quantities drawn from the conditional probabilities will be used to calculate the Bayesian estimates and other related statistics.

Asparouhov and Bengt Muthén (2010b) describe three algorithms for generating $(\mathbf{Z}^{*(t)}, \mathbf{A}^{(t)})$ in Step 3 of (3.5). Method 1 is to put $\mathbf{Z}^{*}$ and $\mathbf{A}$ in one block and draw sample from the following distribution which is separated into two parts

$$p(\mathbf{Z}^{*}, \mathbf{A} \mid **) = p(\mathbf{A} \mid **)\, p(\mathbf{Z}^{*} \mid \mathbf{A}, **) \tag{3.6}$$

where ** represent all the information that $(\mathbf{Z}^{*}, \mathbf{A})$ are conditional on in the Step 3 of (3.5).

In Method 2, suppose there are *D* ordinal variables, then for $d = 1, 2, ...D$, the Gibbs sampler will take D sub-steps as shown in (3.7).

$$\begin{gathered} p(\mathbf{Z}^{*}_{1}, \mathbf{A}_{1} \mid **, \mathbf{Z}^{*}_{d}, \mathbf{A}_{d}, d \neq 1) \\ p(\mathbf{Z}^{*}_{2}, \mathbf{A}_{2} \mid **, \mathbf{Z}^{*}_{d}, \mathbf{A}_{d}, d \neq 2) \\ \vdots \\ p(\mathbf{Z}^{*}_{D}, \mathbf{A}_{D} \mid **, \mathbf{Z}^{*}_{d}, \mathbf{A}_{d}, d \neq D) \end{gathered} \tag{3.7}$$

Each sub-step is again separated into two parts as in (3.6).

Method 3 also takes multiple sub-steps, but generates $\mathbf{Z}^{*}$ and $\mathbf{A}$ separately. It first takes *D* sub-steps for $\mathbf{Z}^{*}$.

$$\begin{array}{c} p(\mathbf{Z}_1^* \mid **, \mathbf{Z}_d^*, \mathbf{A}, d \neq 1) \\ p(\mathbf{Z}_2^* \mid **, \mathbf{Z}_d^*, \mathbf{A}, d \neq 2) \\ \vdots \\ p(\mathbf{Z}_D^* \mid **, \mathbf{Z}_d^*, A, d \neq D) \end{array} \tag{3.8}$$

Then an additional sub-step is carried out for generating all thresholds **A**.

$$p(\mathbf{A} \mid **, \mathbf{Z}^*) \tag{3.9}$$

Generally, Method 1 is most efficient, and Method 3 is least efficient among the three. However, the efficient methods are not always applicable. Method 1 only applies when the variance-covariance matrix of the conditional distribution of $\mathbf{Z}^*$ is a diagonal matrix. If there are non-zero off-diagonal elements in the matrix and no equality constraints are imposed on the thresholds in the model, Method 2 could be used. Method 3 can be used for all situations although it is least efficient.

BSEM with Missing Non-Normal Data

To handle MCAR or MAR data in Bayesian SEM, one additional step is required in the data augmentation and MCMC procedure compared with (3.5). Let $\mathbf{V}_{\text{obs}} = (\mathbf{Y}_{\text{obs}}, \mathbf{Z}^*_{\text{obs}})$ represent the observed continuous data and the latent continuous scores underlying the observed ordinal data, and $\mathbf{V}_{\text{miss}} = (\mathbf{Y}_{\text{miss}}, \mathbf{Z}^*_{\text{miss}})$ represent the missing data. Then (3.5) can be rewritten as

$$\begin{gathered}\text{Step 1} \\ \boldsymbol{\theta}^{(t)} \text{ from } p(\boldsymbol{\theta} \mid \boldsymbol{\Omega}^{(t-1)}, \mathbf{Z}_{\text{obs}}^{*(t-1)}, \mathbf{A}^{(t-1)}, \mathbf{V}_{\text{miss}}^{(t-1)}, \mathbf{Y}_{\text{obs}}, \mathbf{Z}_{\text{obs}}) \\ \text{Step 2} \\ \boldsymbol{\Omega}^{(t)} \text{ from } p(\boldsymbol{\Omega} \mid \boldsymbol{\theta}^{(t)}, \mathbf{Z}_{\text{obs}}^{*(t-1)}, \mathbf{A}^{(t-1)}, \mathbf{V}_{\text{miss}}^{(t-1)}, \mathbf{Y}_{\text{obs}}, \mathbf{Z}_{\text{obs}}) \\ \text{Step 3} \\ \mathbf{V}_{\text{miss}}^{(t)} \text{ from } p(\mathbf{V} \mid \boldsymbol{\theta}^{(t)}, \boldsymbol{\Omega}^{(t)}, \mathbf{Z}_{\text{obs}}^{*(t-1)}, \mathbf{A}^{(t-1)}, \mathbf{Y}_{\text{obs}}, \mathbf{Z}_{\text{obs}}) \\ \text{Step 4} \\ (\mathbf{Z}^{*(t)}, \mathbf{A}^{(t)}) \text{ from } p(\mathbf{Z}^{*}, \mathbf{A} \mid \boldsymbol{\theta}^{(t)}, \boldsymbol{\Omega}^{(t)}, \mathbf{V}_{\text{miss}}^{(t)}, \mathbf{Y}_{\text{obs}}, \mathbf{Z}_{\text{obs}})\end{gathered} \tag{3.10}$$

After convergence, Bayesian estimates and their standard error estimates can be obtained.

In Step 3 of (3.10), the individual missing data points can be separately simulated from their own predictive distributions (see [3.11]). This indicates that missing data patterns have no effect on the simulation procedure (Lee, 2007).

$$p(\mathbf{V}_{\text{miss}}^{(i)} \mid \boldsymbol{\theta}, \boldsymbol{\Omega}, \mathbf{Z}_{\text{obs}}^{*}, \mathbf{A}, \mathbf{Y}_{\text{obs}}, \mathbf{Z}_{\text{obs}}) = \prod_{i=1}^{N} p(\mathbf{v}_{\text{miss}}^{i} \mid \boldsymbol{\theta}, \boldsymbol{\Omega}, \mathbf{z}_{\text{obs}}^{*(i)}, \mathbf{y}_{\text{obs}}^{(i)}, \mathbf{z}_{obs}^{(i)}) \tag{3.11}$$

where $\mathbf{v}_{\text{miss}}^{(i)} = (\mathbf{y}_{\text{miss},}^{(i)} \mathbf{z}_{\text{miss}}^{*(i)})$ is the i^{th} observation in the random sample of size *N*. Also note that in this situation $\mathbf{Z}^{*}$ is not categorized to the corresponding threshold intervals; **A** can be omitted from the conditional distributions in Step 3 of (3.10). Because the MCMC approach draws its sample from the posterior distribution that is conditional on all the other information in the model, MAR data should be handled appropriately.

Performance of BSEM in the Literature

Song and Lee (2002) and Lee and Song (2004a) examined BSEM for normal continuous and categorical MAR data in the contexts of linear and nonlinear structural models. Their studies

indicated that BSEM produced accurate estimates with moderate to large sample sizes (i.e. N = 430 and 1000), and was more accurate and efficient than listwise deletion. Another simulation study conducted by Asparouhov and Muthén (2010a) showed that with N = 1000 the Bayesian estimator produced unbiased estimate for tetrachoric correlation between two binary variables and it was superior to cat-DWLS with MAR data. However, these studies only considered a small number of scenarios. Thus, to better understand the performance of BSEM, a more thorough investigation is warranted.

Chapter 4: Multiple Imputation

Multiple imputation (MI) is a widely used modern missing data technique. Similar to FIML, MI produces unbiased parameter estimates, assuming multivariate normality and MAR (Rubin, 1987; Schafer & Graham, 2002). In fact, the parameter estimates from MI and FIML are expected to be identical if the same hypothesized model and the input data are used, and the number of imputations is infinite (Collins, Schafer, & Kam, 2001; Graham, Olchowski, & Gilreath, 2007; Savalei & Rhemtulla, 2012). However, MI has been found less efficient than FIML (Yuan et al., 2012). In addition, MI is more difficult as it involves multiple phases, while no software package has been created to fully automate them. Despite these disadvantages, there are benefits of using MI. First, it is flexible. A variety of imputation algorithms and imputation models are available that might provide better treatments of non-normal variables and nonparametric relations among the variables (Asparouhov & Muthén, 2010c; White, Royston, & Wood, 2011). Moreover, MI creates complete data sets. Under some circumstances when it is required to use a statistical method that works only with complete data, MI has to be adopted (Enders, 2010; Gottschall, West, & Enders, 2012).

A standard MI procedure involves three phases: 1) imputation phase: generate multiple imputed data with missing values filled in; 2) analysis phase: fit the hypothesized model to each of imputed data sets; and 3) pooling phase: pool the results across imputed data sets to produce the final results. The most commonly used MI method, multivariate normal imputation (MI-MVN), assumes multivariate normality and generates multiple data sets based on a jointly normal distribution. This chapter first introduces the general idea of MI using MI-MVN as an example, and then discusses the other MI methods, latent variable imputation (MI-LV) and

multiple imputation by chained equations (MICE), which vary in the algorithm used in the imputation phase.

The General Idea

Multivariate normal imputation (MI-MVN). The most popular algorithms for MVN imputation are the Gibbs sampler with data augmentation, and the expectation-maximization with bootstrapping.

Gibbs sampler with data augmentation. Similar to the procedure introduced in the BSEM chapter, let $\mathbf{Y}_{\text{obs}}$ and $\mathbf{Y}_{\text{miss}}$ represent the observed and missing data, respectively, and $\boldsymbol{\theta}$ be the imputation parameters which typically contain the elements in a mean vector and covariance matrix under the assumption of multivariate normality. Similar to the procedure described above, the Gibbs sampler starts with initial values on the imputation parameters ($\boldsymbol{\theta}^{(0)}$) and then iterates between two steps. Schafer (2010) describes the two steps as the imputation step (I step) and the posterior step (P step). At the t^{th} iteration, in the I step, missing data values are predicted from the observed data conditional on $\boldsymbol{\theta}^{(t-1)}$ .This is equivalent to drawing random values from the predictive probability distribution of $\mathbf{Y}_{\text{miss}}$, which is typically multivariate normal. In the P step, a random set of imputation parameter values ($\boldsymbol{\theta}^{(t)}$) is drawn from the posterior distribution of the imputation parameters. The two steps at the t^{th} iteration are as follows.

$$\begin{gathered}\text{I step:} \\ \mathbf{Y}_{\text{miss}}^{(t)} \text{ from } p(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t-1)}) \\ \text{P step:} \\ \boldsymbol{\theta}^{(t)} \text{ from } p(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}^{(t)})\end{gathered} \tag{4.1}$$

After the P step, the imputation parameter values are carried to the I step of the next iteration to update the predictive probability distribution of $\mathbf{Y}_{\text{miss}}$. This procedure cycles through many iterations until it reaches convergence. When it converges (after discarding burn-in iterations), the posterior distribution(s) of the imputation parameters stabilizes, the imputed data sets can be then saved. In order to avoid the autocorrelations among the quantities at adjacent iterations, however, the imputed data are saved at every k^{th} iteration rather than every iteration.

Expectation-maximization with bootstrapping. Another algorithm that also assumes multivariate normality is called expectation-maximization with bootstrapping (EMB). This algorithm combines expectation-maximization (EM) algorithm with bootstrapping to obtain the imputation parameter values (Honaker, King, 2010; Honaker, King, & Blackwell, 2011). Specifically, EMB draws *M* bootstrap samples from the original data and then uses the EM algorithm to obtain the maximum likelihood estimates of the mean and the covariance matrix in an imputation model for each bootstrap sample. The EM estimates from each bootstrapped sample is then treated as a random draw of the imputation parameters and used to impute missing data. As a result, *M* imputed data sets are created. For more information on the EM algorithm and bootstrapping, researchers can refer to Dempster et al. (1977) and Efron and Tibshirani (1993), respectively. EMB is theoretically equivalent to the Gibbs sampler and runs much faster than the latter for two reasons: 1) Convergence of the EM algorithm is more straightforward than convergence of Gibbs sampler. It does not require the usual convergence diagnostics in Gibbs sampler; 2) Unlike Gibbs sampler, no autocorrelation exits among the imputation parameter values from different bootstrap samples. Every bootstrap sample is used to generate the imputed datasets (King, Honaker, Joseph, & Scheve, 2001).

The analysis and the pooling phases. After imputation, the target analysis is applied to each imputed data set. The outcomes from the analysis such as point estimates, standard errors, and χ^2 s, are then pooled into the final results. Rubin (1987) created the rules of pooling these quantities. The final point estimates can be obtained by simply taking the average across the imputations (see Equation [4.2]).

$$\bar{\boldsymbol{\theta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{\theta}}_m \tag{4.2}$$

where M indicates the number of imputation data sets, and $\hat{\boldsymbol{\theta}}_m$ represents the parameter estimates for the m^{th} imputation, $m = 1, 2, ...M$.

Pooling the standard errors is not as simple as taking the average. According to Rubin (1987), the multiple imputation standard errors contain two sources of uncertainty: sampling errors had the data been complete and sampling errors due to missing data. Both sources of uncertainty need to be taken into account when pooling the standard errors. Rubin's pooling formula is based on the sampling variances. The final sampling standard errors are just the square root of the pooled sampling variances. For each parameter, the sampling variance contains two parts: the within-imputation variance and the between-imputation variance. The within-imputation variance reflects the sampling variability of complete data, which is the average of the squared standard errors across *M* imputed data sets (see Equation [4.3]).

$$\mathbf{V}_{\mathrm{W}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{SE}_m^2 \tag{4.3}$$

The sampling fluctuation due to missing data is reflected by the variance of the *M* parameter estimates, called between-imputation variance. The formula is given by

$$\mathbf{V}_{\mathrm{B}} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}})^2 \tag{4.4}$$

When the number of imputation data sets goes to infinity, the sampling variability caused by missing data can be sufficiently accounted for by $\mathbf{V}_{\mathrm{B}}$, and thus the total sampling variance is just the sum of $\mathbf{V}_{\mathrm{W}}$ and $\mathbf{V}_{\mathrm{B}}$. In practice, however, only be a finite number of imputed datasets can be created. For this reason, a correction factor needs to be included in the computation of the total sample variance. That is

$$\mathbf{V}_{\mathrm{T}} = \mathbf{V}_{\mathrm{W}} + \mathbf{V}_{\mathrm{B}} + \frac{\mathbf{V}_{\mathrm{B}}}{M} \tag{4.5}$$

where $\frac{\mathbf{V}_{\mathrm{B}}}{M}$ is the correction factor, which decreases as M increases. The pooled standard error is the square root of the total sampling variance.

Obtaining the pooled χ^2 statistic for SEM models is even more complex. Li, Raghunathan, and Rubin (1991) and Meng and Rubin (1992) developed some tools for combining the test statistics across imputations. However, little research has been done to evaluate their performance (Enders, 2010).

Robustness of MI-MVN to non-normality. Demirtas, Freels, and Yucel (2008) did a simulation study to examine the plausibility of the multivariate normality assumption for continuous data. The continuous data were generated based on a broad range of distributions, including normal distribution, t distribution, Laplace distribution, Beta distribution, etc. They suggest that MI-MVN is a reasonable tool for dealing with continuous missing data even when the multivariate normality assumption is violated. One limitation of this study was that most of the non-normal distributions investigated only possessed very mild skewness and kurtosis

(absolute values less than 1). On the other hand, Yuan, et al. (2012) found that MI-MVN could produce biases in both parameter estimates and standard errors with larger levels of kurtosis (around 10). Robustness of MI-MVN under severe non-normality needs to be further studied.

For categorical missing data, it is typically suggested to keep the fractional part of the imputed values resulting from the multivariate normal imputation rather than round them, unless the discrete metric is required by the follow-up analysis (Enders, 2010; Graham, 2009; Graham & Schafer, 1999; Honaker, King, & Blackwell, 2011; Schafer & Graham, 2002).

Latent Variable Imputation (MI-LV)

The latent variable imputation (MI-LV) is specifically used to impute missing ordinal data. The idea is to assume that there is a continuous latent variable underlying each observed ordinal variable (Asparouhov, & Muthén, 2010c). The underlying latent variables are typically assumed to follow a multivariate normal distribution. The imputed values are first imputed at the latent variable level using a normal data model and then discretized based on estimated thresholds. This latent variable model is a formulation of a cumulative/ordinal probit model (Cowles, 1996).

Let $\mathbf{Y}^*$ represent the vector of latent variables underlying a vector of categorical variables. Equation (4.6) shows the default imputation model for $\mathbf{Y}^*$, which is a saturated model.

$$\mathbf{Y}^* \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4.6}$$

where *MVN* stands for multivariate normal distribution; $\boldsymbol{\mu}$ is a vector of latent means, which are usually fixed to 0 for identification purpose, so that all of the thresholds can be freely estimated; and $\boldsymbol{\Sigma}$ is the covariance matrix of the latent variables. The diagonal elements of $\boldsymbol{\Sigma}$ are fixed at 1 to set the scale and the off-diagonal elements are freely estimated. Specifically, for complete

cases, the **Y*** values follow a truncated normal distribution, such that they are bounded by the appropriate threshold parameters. For incomplete cases, the **Y*** values are unbounded since they are not able to condition on the discrete scores.

Once the imputed values are obtained for **Y***, they are discretized using the thresholds as follows.

$$\mathbf{z}_j = c \Leftrightarrow a_{(c\text{-}1)j} < \mathbf{z}_j^* < a_{cj} \tag{4.7}$$

where c is an integral value that indicates the category, $c = 1, 2, \ldots, C$, and a_{cj} is the c^{th} threshold for the j^{th} variable ($a_{0j} = -\infty$, and $a_{Cj} = +\infty$).

Because missing data are imputed at the latent variable level and the latent variables are assumed to follow a multivariate normal distribution, it is straightforward to establish the joint distribution even when there is a mixture of ordinal variables and continuous variables. Using the Gibbs sampler, for example, the imputation process involves two more steps than MI-MVN shown in (4.1), which are drawing thresholds parameters and latent variable scores from the joint distribution, and categorizing latent normal imputations using the threshold parameters (Cowles, 1996).

Multiple Imputation by Chained Equations (MICE)

Another popular imputation algorithm is termed multiple imputation by chained equations (MICE; van Buuren et al., 2006; van Buuren & Groothuis-Oudshoorn, 2011). This algorithm is also known as fully conditional specification or sequential regression multivariate imputation. Different than MI-MVN or MI-LV, MICE does not impute based on a joint distribution. Rather, it imputes missing data on a variable-by-variable basis. Prediction of missing data on each variable is conditional on the current values of the other variables at a

specific iteration. The imputation model for each missing data variable can be specified individually. Thus this algorithm is very flexible in accommodating missing data with different scales.

Parametric MICE. In MICE, the Gibbs sampler takes the I step and the P step to draw parameter values and impute missing data for every incomplete variable. Let $\mathbf{y}_1, \mathbf{y}_2, ... \mathbf{y}_p$ be the p variables that need to be imputed, and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ... \boldsymbol{\theta}_p$ be parameters that describe the distribution of each variable, respectively. The two steps for every variable at the t^{th} iteration can be described as follows.

$$\begin{aligned}
\boldsymbol{\theta}_1^{(t)} &\text{ from } p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1^{\text{obs}}, \mathbf{y}_2^{(t-1)}, ..., \mathbf{y}_p^{(t-1)}) \\
\mathbf{y}_1^{\text{miss}(t)} &\text{ from } p(\mathbf{y}_1^{\text{miss}} \mid \mathbf{y}_1^{\text{obs}}, \mathbf{y}_2^{(t-1)}, ..., \mathbf{y}_p^{(t-1)}, \boldsymbol{\theta}_1^{(t)}) \\
&\vdots \\
\boldsymbol{\theta}_p^{(t)} &\text{ from } p(\boldsymbol{\theta}_p \mid \mathbf{y}_p^{\text{obs}}, \mathbf{y}_1^{(t)}, \mathbf{y}_2^{(t)}, ..., \mathbf{y}_p^{(t)}) \\
\mathbf{y}_p^{\text{miss}(t)} &\text{ from } p(\mathbf{y}_p^{\text{miss}} \mid \mathbf{y}_p^{\text{obs}}, \mathbf{y}_1^{(t)}, \mathbf{y}_2^{(t)}, ..., \mathbf{y}_p^{(t)}, \boldsymbol{\theta}_p^{(t)})
\end{aligned} \tag{4.8}$$

One major advantage of the MICE algorithm is that it does not require any joint distribution and the imputation model can be easily tailored to the nature of each variable. When a variable **y** is normal, **θ** may contain linear regression coefficients or elements in mean vector and covariance matrix; when **y** is binary, **θ** may contain logistic regression coefficients; when **y** is ordinal, **θ** may contain ordinal logistic regression coefficients, etc. Correspondingly, the predictive probability distribution of $\mathbf{y}^{\text{miss}}$ is tailored to the scale. That is, normal data are drawn from a normal distribution; binary data are drawn from a binomial distribution; and ordinal data are drawn from a cumulative logistic distribution, etc. Research showed that MICE usually only requires a small number of iterations (e.g., 5 to 20) to converge (e.g., van Buuren & Groothuis-Oudshoorn, 2011; van Buuren, 2012).

The MICE algorithm provides a general framework to implement various imputation models, including not only parametric models, such as linear regression and cumulative logistic regression, but also semi-parameter or non-parametric models. The main benefit of using semi-parameter or non-parametric models is that they rely less on distributional assumptions and can preserve the original scales and the non-parametric relations in the data. In the following two sections, two other MICE methods, i.e., MICE with a semi-parametric technique called predictive mean matching (PMM; Little, 1998) and MICE with a nonparametric technique called random forest (RF; Breiman, 2001) are introduced in details.

MICE with predictive mean matching (MICE-PMM). The idea of MICE with predictive mean matching (MICE-PMM) is to impute each missing value by randomly drawing a value from its nearest observed neighbors (candidate donors) on the same variable. PMM is a semi-parametric imputation approach. It does not require a specific model to define the distribution of missing data; however, a parameter predictive model (usually a linear regression model) is needed to determine the candidate donor pool (Heitjan & Little, 1991; Schenker & Taylor, 1996).

To illustrate, let $\mathbf{Y}$ be a data matrix with *p* variables; $\mathbf{y}_j = (\mathbf{y}_j^{\text{obs}}, \mathbf{y}_j^{\text{miss}})$, for *j* = 1, 2,.., *p*, where $\mathbf{y}_j^{\text{obs}}$ represents the observed data, and $\mathbf{y}_j^{\text{miss}}$ is the missing data; and $\dot{\mathbf{Y}}$ is the currently imputed data matrix. A typical PMM approach implemented in MICE involves four steps (van Buuren, 2012):

1. For each $\mathbf{y}_j$, *j* = 1, 2, ..., *p* , fill in initial imputations sequentially by random draws from $\mathbf{y}_j^{\text{obs}}$. This results in a complete data matrix $\dot{\mathbf{Y}}^{(0)}$.
2. For each $\mathbf{y}_j$, *j* = 1, 2, ..., *p* , update $\dot{\mathbf{Y}}^{(0)}$ as follows:

a. Estimate a parametric model (e.g., linear regression) based on $\dot{\mathbf{Y}}$ and obtain the predicted values, $\hat{\mathbf{y}}_j = (\hat{\mathbf{y}}_j^{\text{obs}}, \hat{\mathbf{y}}_j^{\text{miss}})$.

b. For each element in $\mathbf{y}_j^{\text{miss}}$, match it to a certain number of values in $\mathbf{y}_j^{\text{obs}}$ (candidate donors) according to the distances between their predictive values, i.e., $|\hat{\mathbf{y}}_{jk}^{obs} - \hat{\mathbf{y}}_{jl}^{miss}|$, where *k* represents the k^{th} observation in $\hat{\mathbf{y}}_j^{\text{obs}}$, and *l* denotes the *l*th observation in $\hat{\mathbf{y}}_j^{\text{miss}}$.

c. Randomly select one donor as the imputed value for each element in $\mathbf{y}_j^{\text{miss}}$.

3. Repeat step 2 for *T* times (*T* iterations), yielding one imputed data set.
4. Repeat step 3 for *M* times, yielding *M* imputed sets.

Multiple versions of PMM have been developed by varying one or some of the computational details in the above steps. First, in addition to the liner regression model, other parametric models can be also used to predict the means of the data (e.g., Di Zio & Guarnera, 2009). Second, the traditional way to estimate the parameters in a parametric model ignores the sampling variability of the parameters (van Buuren, 2012). It could be problematic, especially when there are only a small number of predictors (Heitjan & Little, 1991). This problem can be alleviated by using the Bayesian approach (i.e., randomly drawing parameter values from its posterior distribution) or the bootstrapping approach (Koller-Meinfelder, 2009). Third, there are many ways to determine the number of candidate donors (denoted as *d*; see Andridge & Little, 2010, for more details). Generally speaking, $d = 1$ is not a good option because it may create too little variability across the imputed data sets. On the other hand, a high *d* value may also cause problems, as a high *d* may increase the likelihood of bad matches. Common values for *d* are 3, 5

and 10 (van Buuren, 2012; Morris, White, & Royston, 2014), however, further research is needed to establish a guideline for specifying *d*.

PMM in combination with MICE could be more robust than a parametric approach, because it relaxes the distribution assumption of the parametric imputation (Di Zio and Guarnera, 2009; Morris et al., 2014). However, it also has obvious limitations. As noted by van Buuren (2012, p. 74), "It cannot be used to extrapolate beyond the range of the data, or interpolated within the range of data if the data at the interior are sparse. Also, it may not perform well with small datasets."

MICE with random forests (MICE-RF). Random forests (RF) refer to one of non-parametric recursive partitioning methods for regression and classification. Different than classical regression and classification methods, recursive partitioning predicts a response variable by successively splitting the data set based on one predictor at a time so that the subsets become more homogeneous with each split (Breiman, 2001). Since this splitting procedure and resulting subsets can be represented by a tree structure, these methods are also called decision tree methods (James, Witten, Hastie & Tibshirani, 2013).

The simplest recursive partitioning method is termed classification and regression trees (CART), which is referred to as classification trees or regression trees depending on whether the response variable is continuous or categorical. A regression tree is illustrated using the Hitters example (James, et al, 2013). In this example, the salary of a baseball player (**y**) is predicted by the number of years that he has played in the major leagues ($\mathbf{x}_{\mathbf{years}}$) and the number of hits that he made in the previous years ($\mathbf{x}_{\mathbf{hits}}$). The whole data set can be first split into two subsets based on $\mathbf{x}_{\mathbf{year}}$ values. Players with $\mathbf{x}_{\mathbf{year}} < 4.5$ are assigned to the left branch of the tree (denoted as R1), and those with $\mathbf{x}_{\mathbf{year}} \geq 4.5$ go to the right branch. Next, the right branch can be further split based

on $\mathbf{x}_{\mathbf{hits}}$. Players with $\mathbf{x}_{\mathbf{year}} \geq 4.5$ and $\mathbf{x}_{\mathbf{hits}} < 117.5$ go to one sub-branch (denoted as R2) and those with $\mathbf{x}_{\mathbf{year}} \geq 4.5$ and $\mathbf{x}_{\mathbf{hits}} \geq 117.5$ go to the other (denoted as R3). The regions of R1, R2 and R3 are known as terminal nodes or leaves of the tree. For the observations within the same region, the predicted values are all equal to the mean of their response values.

There are multiple possible ways to build the regression tree (i.e., construct R1, R2 and R3) depending on different specifications of the order of the predictors and cut points. One popular criteria to determine where to make a split is to minimize the residual sum of squares (*RSS*), given by

$$RSS = \sum_{j=1}^{J} \sum_{i=R_j} (\mathbf{y}_i - \hat{\mathbf{y}}_{R_j})^2 \tag{4.9}$$

where $\hat{\mathbf{y}}_{R_j}$ is the mean of response values within the region R_j, j = 1, 2,.., J, and i indicates i^{th} observation in each R_j. The tree grows until a stopping criterion is reached. A classification tree is based on the same idea as a regression tree, except that the response variable is categorical. One popular criteria to evaluate a split in a classification tree is Gini-index (James, et al, 2013), which is

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4.10}$$

where $\hat{p}_{mk}$ indicates the proportion of the observations in the m^{th} region that are from the k^{th} category. More details on splitting and pruning regression and classification trees can be found in James, et al. (2013).

CART could be very unstable, because it creates only a single tree and the splitting procedure highly depends on the distribution of observations in the sample (Strobl, Malley, &

Tutz, 2009; Doove, van Buuren, & Dusseldorp, 2014). One solution to this problem is termed random forests (RF). In RF, the algorithm to grow a single tree is the same as in CART; the difference is that RF generates multiple samples based on the original data and creates a tree for each sample. Either of the following two ways can be used to create multiple samples: 1) bootstrapping and 2) bootstrapping combined with random selections of a smaller group of predictors. By averaging the results across multiple trees, RF takes into account the variation of bootstrapped samples and results in a more stable solution.

Some work has been done to combine RF with single imputation methods (e.g., Iacus & Porro, 2006; Stekhoven & Bühlmann, 2012). The major problem with these methods was that they failed to account for the sampling variability due to missing data when analyzing the single imputed data (Shah, Bartlett, Carpenter, Nicholas, & Hemingway, 2014; Doove et al., 2014). This problem can be overcome by incorporating RF with MICE.

When implementing RF with MICE, the missing values (either continuous or categorical) on each variable are imputed based on their RF predicted values. Because RF does not rely on distributional assumptions or parametric models, it has a potential to accommodate non-normal missing data and non-linear relationships (Shah et al., 2014; Doove et al, 2014). Doove et al (2014) have described an algorithm to implement RF in MICE, which involves four steps. Suppose a data matrix $\mathbf{Y}$ contains p variables; $\mathbf{y}_j = (\mathbf{y}_j^{obs}, \mathbf{y}_j^{miss})$, for j = 1, 2,.., p, where $\mathbf{y}_j^{obs}$ represents the observed data, and $\mathbf{y}_j^{miss}$ is the missing data; and $\dot{\mathbf{Y}}$ is the currently imputed data matrix, then

1. For each $\mathbf{y}_j$, j = 1, 2, ..., p, fill in initial imputations by random draws from $\mathbf{y}_j^{obs}$ sequentially. This results in a complete data matrix $\dot{\mathbf{Y}}^{(0)}$.

2. For each $\mathbf{y}_j$, j = 1, 2, ..., p, update $\dot{\mathbf{Y}}^{(0)}$ as follows:

    a. Draw b bootstrap samples from $\dot{\mathbf{Y}}$, only for observations in $\mathbf{y}_j^{obs}$.

    b. Fit one tree on every bootstrap sample drawn in step 2a, with the best splits decided either with or without a random selection of a smaller group of predictors. This results in b trees; each has several leaves, and each leaf contains a subset of elements of $\mathbf{y}_j^{obs}$. The values in the subset are called donors.

    c. For each element of $\mathbf{y}_j^{\text{miss}}$, determine in which leaf it will end up according to the b trees fitted in step 2b. This results in b leaves with donors for each element of $\mathbf{y}_j^{\text{miss}}$.

    d. For each element of $\mathbf{y}_j^{\text{miss}}$, take all donors in the b leaves from step 2c and randomly select one to replace the initial imputation.

3. Repeat step 2 for T times (T iterations), yielding one imputed data set.

4. Repeat step 3 for M times to create M imputed data sets.

Performance of the Imputation Methods in the Literature

MI-LV has received increasing attention in recent years. Asparouhov and Muthén (2010c) compared the performance of MI-LV followed by cat-DWLS to direct cat-DWLS in estimating a growth model of 5 binary variables observed at 5 time points. They concluded that this imputation method outperformed direct cat-DWLS under MAR by providing more accurate parameter estimates and higher confidence interval coverage. Wu, Jia, and Enders (2015) found that MI-LV performed very well in a scenario where the ordinal variables are to be aggregated to scale scores for regression analysis, regardless of the missing data proportions, sample sizes,

asymmetry of categories, and numbers of categories. However, the performance of MI-LV in SEM has not been systematically examined.

Research on the performance of MICE with non-normal data typically focuses on manifest variable models. In terms of MICE with parametric imputation models, van Buuren et al. (2006) found that MICE in combination with logistic regression (MICE-logit) was superior to listwise deletion when estimating odds ratio. Van Buuren (2007) recommended using MICE-logit rather than MI-MVN for ordinal logistic regression analysis. Wu et al. (2015), however, found that using MICE-logit to impute ordinal missing values could lead to large bias in estimating reliability coefficients, mean scale scores, and regression coefficients of predicting one scale score from another, especially when sample size was small, item distributions were asymmetric, and the number of categories was more than five.

Regarding semi-parametric or nonparametric imputation models, MICE-PMM was found to outperform listwise deletion, single imputation and MICE-logit in fitting a Cox proportional hazards model with moderately skewed data (Marshall, Altman, & Holder, 2010). Doove et al (2014) compared MICE-RF (with and without selection of a small group of predictors) with MICE-logit under a scenario where data were categorical under MAR and the analysis models were logistic regression models with interaction effects. The findings suggest that MICE-RF produced more accurate estimate of the interaction effect and was more efficient than MICE-logit. Little research was found in the literature to examine how MICE-RF performs with missing non-normal continuous data. None of the methods has been examined in the context of SEM.

Chapter 5: Study I - Non-Normal Continuous Data

Research Questions

This study focuses on five methods for dealing with non-normal continuous missing data in SEM, including robust FIML (RFIML), normal-theory-based Bayesian SEM (BSEM), multivariate normal imputation (MI-MVN), MICE with predictive mean matching (MICE-PMM), and MICE with random forests (MICE-RF). Table 1 demonstrates the characteristics of the five methods in terms of their non-normality strategy (rescaling vs. Bayesian method), missing data strategy (direct method vs. imputation method), distributional assumption and software implementation.

Table 1. Summary of Methods for Missing Non-Normal Continuous Data

Method Label	Method Description	Strategy for Non-Normality	Strategy for Missing Data	Normality Assumption	Software
RFIML	FIML with the rescaling strategy proposed by Yuan and Bentler (2000).	Rescaling	Direct Method	No	R
MI-MVN	Multivariate normal imputation using EMB algorithm (Honaker, King, & Blackwell, 2011). Robust ML (RML; Satorra & Bentler, 1994) is used in the analysis phase.	Rescaling	Imputation Method	Yes (in imputation phase)	R
MICE-PMM	MICE with predictive mean matching using the Bayesian approach (van Buuren, 2012). Robust ML (RML; Satorra & Bentler, 1994) is used in the analysis phase.	Rescaling	Imputation Method	No	R
MICE-RF	MICE with random forests, which involves a random selection of a smaller group of predictors at each split (Doove et al., 2014). Robust ML (RML; Satorra & Bentler, 1994) is used in the analysis phase.	Rescaling	Imputation Method	No	R
BSEM	Gibbs sampler with data augmentation (Lee, 2007; Asparouhov, & Muthén, 2010b).	Bayesian Method	Direct Method	Yes	Mplus

Three research questions were addressed in this study.

Question 1: To what extent are the normal-theory-based MI-MVN and BSEM robust to non-normal continuous data?

Question 2: How are the methods influenced by sample size, degree of non-normality, missing data mechanism and missing data proportion?

Question 3: Which method performs best under a variety of conditions, with respect to sample size, degree of non-normality, missing data mechanism, and missing data proportion?

Method

Data generation model. Data were generated based on the structural equation model used in Enders (2001b, see Figure 1). The structure of this model is also commonly seen in the SEM literature (e.g., Bollen, 1989; Palomo, Dunson, & Bollen, 2011). The model involved three latent variables: $\boldsymbol{\eta 1}$, $\boldsymbol{\eta 2}$ and $\boldsymbol{\eta 3}$. $\boldsymbol{\eta 3}$ was predicted by $\boldsymbol{\eta 1}$ and $\boldsymbol{\eta 2}$. As shown in Figure 1, the values of the structural paths among the three variables were 0.4 ($\boldsymbol{\eta 2}$ regressed on $\boldsymbol{\eta 1}$), 0.286 ($\boldsymbol{\eta 3}$ regressed on $\boldsymbol{\eta 2}$) and 0.286 ($\boldsymbol{\eta 3}$ regressed on $\boldsymbol{\eta 1}$). The variance of $\boldsymbol{\eta 1}$ was fixed to 1 for identification purpose. The residual variances of $\boldsymbol{\eta 2}$ and $\boldsymbol{\eta 3}$ were set to 0.84 and 0.771 so that their variances were both equal to 1. Each latent variable was indicated by three manifest variables with all loadings set to 0.70. The residual variances on the indicators were all set to 0.51.

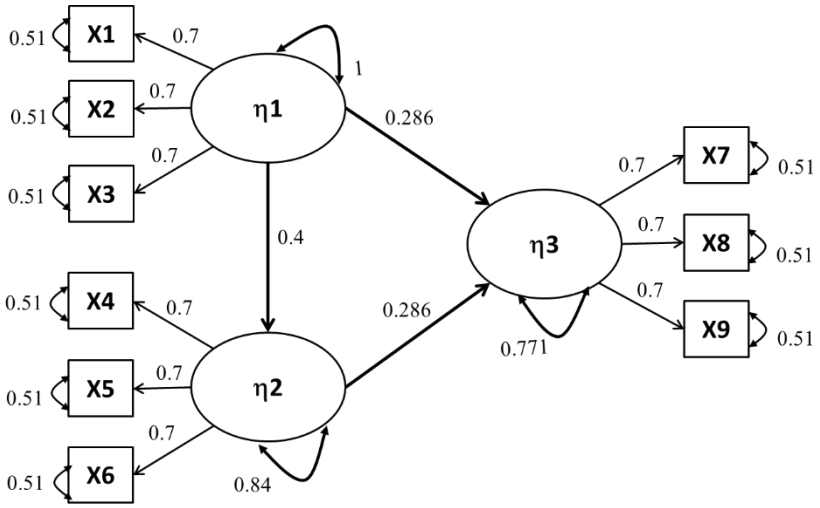

Figure 1. Structural equation model for data generation.

Design factors.

Degrees of non-normality. The non-normal continuous data were generated following the method proposed in Vale and Maurelli (1983) and Fleishman (1978). The levels of non-normality were specified by three combinations of univariate skewness (*S*) and excess kurtosis (*K*): mild non-normality (*S* = 1.5, *K* = 3), moderate non-normality (*S* = 2, *K* = 7), and severe non-normality (*S* = 3, *K* = 21). The corresponding approximate multivariate kurtoses (Mardia, 1970) were 143, 187, and 314, respectively. These levels of non-normality were reflective to the data observed in applied research (Curran et al, 1996), and were similar to those used in Enders (2010b) and Savalei and Falk (2014). All manifest variables had the same degree of non-normality under each condition. The correlation matrix of the non-normal data was comparable across the different levels of non-normality and was comparable to that of the normal data.

Sample size. Sample size was manipulated at two levels: small (300) and large (600).

Missing data mechanism. Missing data were created on two of the indicators for each latent variable. Specifically, missing values occurred on **x1**, **x2**, **x4**, **x5**, **x7** and **x8**. The missing

data were generated using three mechanisms: MCAR, MAR-Head, and MAR-Tail (described below). MCAR data were generated by randomly deleting a desired proportion of values on each missing data variable.

To generate MAR data, the rank order of the values on each of the fully-observed variables (**x3**, **x6** and **x9**) was used to determine the probability of having a missing observation on the other two manifest variables for the same latent variable. For example, the missingness on **x1** and **x2** was determined by **x3**. MAR-Head data were generated based on the rank in an ascending order. The probability of having missing data on **x1** was computed as 1 minus the ascending order of the value on **x3**. Because all variables were positively correlated, the smaller the value on **x1** was, the higher probability of missingness it had. In addition, because all manifest variables were also positively skewed, this mechanism led to more missing data on the head of the distribution. MAR-Tail data were generated in a similar way except that the probability of being missing is determined by the descending order of the values on the missing data determiners. Under MAR-Tail, there were more missing data on the tail of the distribution.

Missing data proportion. Missing data proportion was manipulated at two levels: small (15%) and large (30%).

One thousand replicated samples were created in each of the fully-crossed conditions ($3 \times 2 \times 3 \times 2 = 36$). In order to differentiate the effect of non-normality and missingness, evaluation was also conducted on the direct methods for complete non-normal data (i.e., robust ML and normal-BSEM). Six (3×2) more conditions were therefore added in the study. The analysis model was the same as the data generation model. For the imputation methods, 50 imputed data sets were generated following the guideline of White et al. (2011). Figure 2 demonstrates the

distributions of one manifest variable from one replication with various degrees of non-normality and missing data mechanisms.

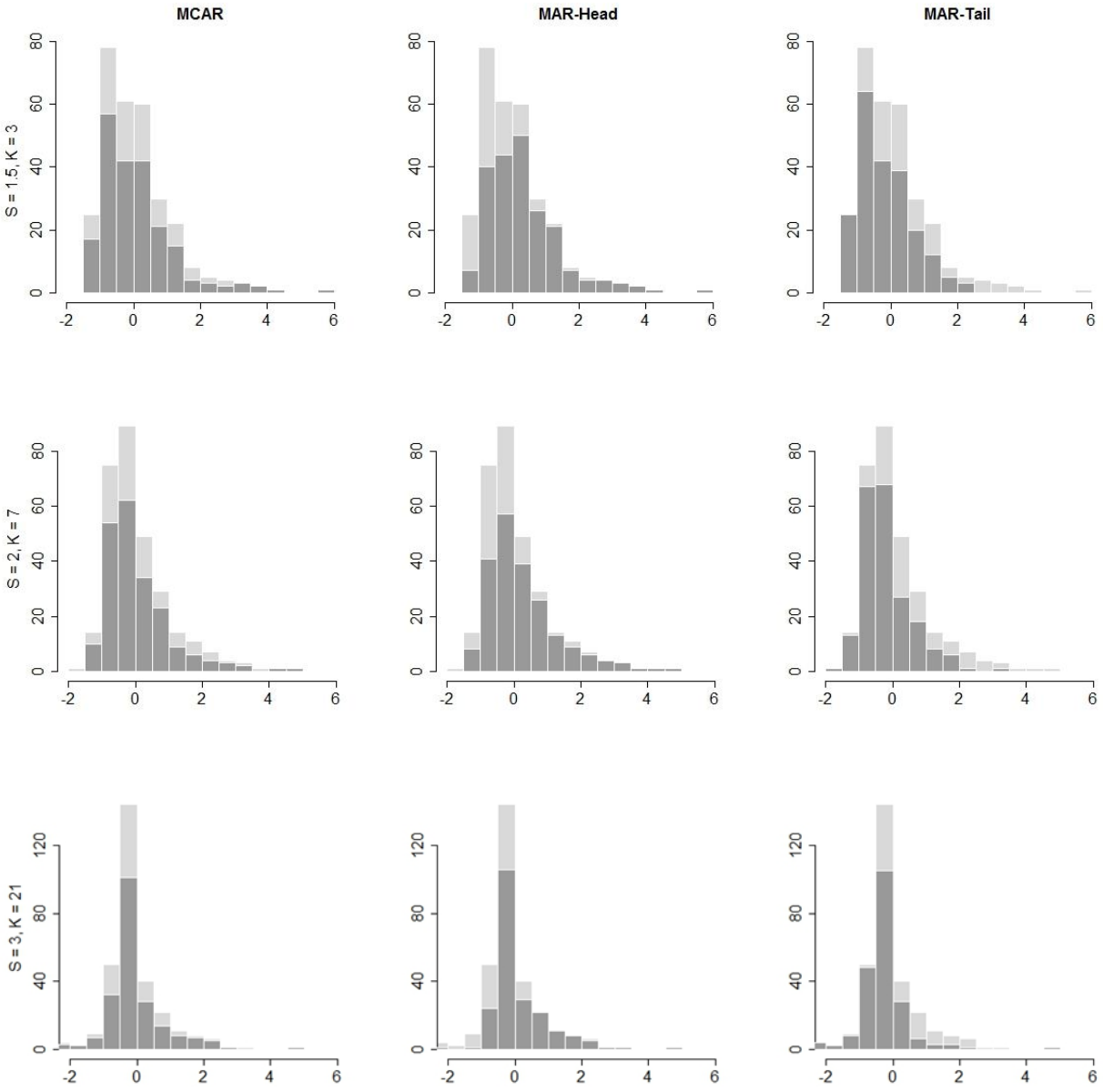

Figure 2. Distributions of x1 (continuous) for one replication with N = 300 before (light grey) and after (dark grey) imposing 30% missing data.

Computational characteristics. Data were generated in R (R Core Team, 2014) using the function gen.nonnormal(·) provided by Zopluoglu (2013) . RFIML was implemented in lavaan (Rosseel, 2012), in which the convergence threshold (relative tolerance) is set at 10^{-10}. Normal-BSEM was conducted via Mplus 7.0 (Muthén & Muthén, 2008-2012) with the default non-informative priors [i.e. the prior for any of the loadings and regression paths is a normal distribution, $N(0,\infty)$; the prior for any of the variances is an inverse Gamma distribution, $IG(-1, 0)$] (Asparouhov & Muthén, 2010a). Two MCMC chains were used in normal-BSEM and the convergence was achieved when Proportional Scale Reduction (PSR) based on the last half of the iterations is small enough, i.e., between 1.05 and 1.1 for all parameters (Asparouhov & Muthén, 2010b).

MI-MVN was implemented using the R package Amelia (Honaker, King, & Blackwell, 2011). The convergence threshold for EM was equal to 10^{-4}. The MICE methods were implemented through functions in the R package mice (van Buuren & Groothuis-Oudshoorn, 2011) with 10 burn-in iterations (van Buuren, Groothuis-Oudshoorn, & Rubin, 2006; White, et al., 2011). Number of donors for MICE-PMM is set to 5 according to Morris et al (2014). For MICE-RF, a minimum leaf size of 5 or 1 was used to create regression trees or classification trees, respectively (Liaw & Wiener, 2002). The number of bootstrap samples (i.e., the number of trees) was set to 10 (Doove et al, 2014). Lavaan was used to analyze all the imputed data with the robust ML (RML) estimator. For all MI methods, a replication reached convergence when the model converged with all of its imputed data sets.

Outcome measures. The performance of the five methods was evaluated on four outcomes: convergence failures and outliers, relative bias in parameter estimates, mean squared error (MSE), and confidence interval (or credible interval) coverage (CIC).

Convergence failures and outliers. The number of convergence failures was the count of non-converged replications in each design cell. In addition, outliers were removed before computing the other three outcomes. An outlier was defined as a replication that converged to a solution with at least one extreme value of parameter or standard error estimate. The outliers were identified and removed using the following rules. First, replications with a standard error estimate greater than 10 were removed. Second, after the first step, a replication was treated as an outlier if at least one of its parameter estimates was 10 standard deviations away from the mean value in the design cell. Third, for each replication, if at least one of its standard error estimates was 10 standard deviations away from the mean value in the design cell, the replication was treated as an outlier and removed.

Relative bias in parameter estimates. The relative bias in parameter estimates is the percentage of the raw bias relative to the true population value.

$$Est\ Bias = \frac{(\bar{\theta}_{est} - \theta_0)}{\theta_0} \times 100\% \tag{5.1}$$

where the numerator represents the raw bias, which is the difference between the average parameter estimate across replications within a design cell ($\bar{\theta}_{est}$) and the population value (θ_0). According to Hoogland and Boomsma (1998), a relative bias that is less than 5% would be considered acceptable. However, Muthén, Kaplan and Hollis (1987) believe that "a bias of less than 10 - 15% may not be serious in most SEM contexts".

Mean squared error (MSE). MSE is defined as the expected squared distance between the parameter estimates and true population value. It can also be decomposed into the sum of the squared bias and the variance of the estimate. Of the two components, the first measures the

accuracy of the method and the second measures the precision of the estimator. A better estimator should yield a smaller MSE.

Confidence interval (or credible interval) coverage (CIC). Confidence interval coverage is estimated as the percentage of replications in a design cell that lead to 95% confidence intervals containing the population value. In the Bayesian framework, the 95% credible interval coverage is used instead. Ideally, the CIC values should be equal to 95%. Following Collins et al. (2001), a coverage value below 90% is considered problematic.

Results

The results of the two direct complete data methods are summarized in

Table 2. For the complete data, RML and normal-BSEM produced very small biases (1% - 3%), and the MSEs from the two methods were comparable across sample sizes and missing data mechanisms. The CICs from normal-BSEM were lower than those from RML, and strongly affected by the degree of non-normality. Specifically, under severe non-normality, the CICs from normal-BSEM fell below 90%. The results for missing non-normal continuous data are shown in Table 3 - Table 5, and the patterns in the outcomes are described as follows.

Table 2. Results for Complete Non-Normal Data

	Mild Non-Normality					Moderate Non-Normality					Severe Non-Normality				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
							N = 300								
RML	0	0	1%	0.010	94%	0	0	1%	0.012	94%	0	0	2%	0.019	91%
BSEM	0	0	1%	0.010	92%	0	0	2%	0.013	90%	0	0	3%	0.020	82%
							N = 600								
RML	0	0	1%	0.005	94%	0	0	1%	0.006	95%	0	0	1%	0.009	93%
BSEM	0	0	1%	0.005	92%	0	1	1%	0.006	90%	0	0	1%	0.009	82%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RML's MSE under the same condition, or CIC < 90%.

Convergence failures and outliers. The convergence rates in all design cells were very high (97% - 100%), and were not affected by the design factors. Among the five methods, only MICE-PMM tended to produce a slightly higher number of convergence failures (27) under N = 300 and 30% MAR-Tail missingness. Only a few replications (37 out of 36000 replications) were identified as outliers.

Table 3. Results for Missing Mildly Non-Normal Data

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
							N = 300, Mprop = 15%								
RFIML	0	0	1%	0.012	94%	0	0	3%	0.013	94%	0	0	-3%	0.012	93%
MI-MVN	0	0	1%	0.012	92%	0	0	3%	0.013	94%	0	0	-2%	0.013	92%
MICE-PMM	0	0	1%	0.012	93%	0	0	3%	0.013	93%	0	0	1%	0.016	94%
MICE-RF	0	0	3%	0.012	94%	0	0	3%	0.013	94%	0	0	14%	0.022	98%
BSEM	0	0	2%	0.012	92%	0	0	4%	0.014	93%	0	0	0%	0.014	93%
							N = 300, Mprop = 30%								
RFIML	0	0	1%	0.013	94%	0	0	5%	0.016	95%	0	1	-4%	0.016	91%
MI-MVN	1	1	1%	0.014	93%	0	0	4%	0.017	94%	2	2	-1%	0.017	93%
MICE-PMM	0	0	1%	0.014	94%	0	0	4%	0.016	93%	27	27	2%	0.022	94%
MICE-RF	0	0	7%	0.015	96%	0	0	7%	0.018	95%	6	7	23%	0.040	99%
BSEM	0	0	3%	0.015	93%	0	2	8%	0.019	95%	0	5	3%	0.020	92%
							N = 600, Mprop = 15%								
RFIML	0	0	1%	0.006	94%	0	0	2%	0.006	95%	0	0	-4%	0.006	93%
MI-MVN	0	0	1%	0.006	93%	0	0	2%	0.006	94%	0	0	-3%	0.006	93%
MICE-PMM	0	0	1%	0.006	94%	0	0	3%	0.006	94%	0	0	-1%	0.008	92%
MICE-RF	0	0	2%	0.006	95%	0	0	3%	0.006	95%	0	0	12%	0.011	96%
BSEM	0	0	1%	0.006	92%	0	0	3%	0.006	93%	0	0	-3%	0.006	91%
							N = 600, Mprop = 30%								
RFIML	0	0	1%	0.007	94%	0	0	4%	0.008	95%	0	0	-6%	0.008	91%
MI-MVN	0	0	1%	0.007	93%	0	0	4%	0.008	94%	0	0	-3%	0.008	92%
MICE-PMM	0	0	1%	0.007	94%	0	0	5%	0.008	94%	0	0	-1%	0.012	92%
MICE-RF	0	0	5%	0.007	96%	0	0	6%	0.008	95%	0	0	18%	0.020	97%
BSEM	0	0	2%	0.007	92%	0	0	6%	0.008	94%	0	0	-3%	0.008	90%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%.

Table 4. Results for Missing Moderately Non-Normal Data

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	0	0	2%	0.014	94%	0	0	3%	0.016	93%	0	0	-6%	0.014	90%
MI-MVN	0	0	2%	0.014	92%	0	0	3%	0.016	92%	0	0	-4%	0.014	91%
MICE-PMM	0	0	2%	0.014	93%	0	0	3%	0.016	92%	0	0	1%	0.018	93%
MICE-RF	0	0	4%	0.014	94%	0	0	3%	0.016	93%	0	0	16%	0.026	98%
BSEM	0	0	3%	0.015	90%	0	0	4%	0.017	91%	0	1	-3%	0.015	89%
N = 300, Mprop = 30%															
RFIML	0	0	2%	0.017	93%	0	0	6%	0.021	94%	0	0	-7%	0.018	88%
MI-MVN	0	0	2%	0.018	92%	0	0	5%	0.022	93%	2	2	-3%	0.017	91%
MICE-PMM	0	0	3%	0.018	93%	0	0	5%	0.020	92%	15	16	3%	0.025	92%
MICE-RF	0	0	8%	0.019	96%	0	0	7%	0.021	94%	3	5	24%	0.043	99%
BSEM	0	1	4%	0.019	91%	0	4	9%	0.026	92%	0	2	0%	0.020	89%
N = 600, Mprop = 15%															
RFIML	0	0	1%	0.007	94%	0	0	2%	0.008	95%	0	0	-7%	0.007	90%
MI-MVN	0	0	1%	0.007	93%	0	0	2%	0.008	94%	0	0	-5%	0.007	91%
MICE-PMM	0	0	1%	0.007	94%	0	0	3%	0.008	94%	0	0	0%	0.010	92%
MICE-RF	0	0	3%	0.007	95%	0	0	3%	0.008	95%	0	0	14%	0.013	97%
BSEM	0	0	1%	0.007	90%	0	0	3%	0.008	90%	0	0	-6%	0.007	86%
N = 600, Mprop = 30%															
RFIML	0	0	1%	0.008	94%	0	0	5%	0.010	95%	0	0	-9%	0.009	88%
MI-MVN	0	0	2%	0.009	92%	0	0	4%	0.010	94%	0	0	-5%	0.008	89%
MICE-PMM	0	0	2%	0.008	94%	0	0	5%	0.010	93%	0	0	0%	0.013	91%
MICE-RF	0	0	6%	0.009	95%	0	0	6%	0.010	95%	0	0	19%	0.021	98%
BSEM	0	0	2%	0.009	90%	0	0	6%	0.011	91%	0	0	-6%	0.009	85%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%.

Table 5. Results for Missing Severely Non-Normal Data

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	0	1	3%	0.022	91%	0	0	3%	0.025	90%	0	1	-7%	0.020	86%
MI-MVN	0	1	3%	0.022	88%	0	0	3%	0.025	88%	1	1	-4%	0.019	88%
MICE-PMM	0	0	3%	0.021	90%	0	0	4%	0.025	88%	3	3	8%	0.030	93%
MICE-RF	0	0	6%	0.022	92%	0	0	5%	0.026	90%	0	0	23%	0.042	98%
BSEM	0	0	4%	0.023	83%	0	0	5%	0.027	82%	0	0	-3%	0.022	83%
N = 300, Mprop = 30%															
RFIML	0	0	4%	0.027	89%	0	0	6%	0.034	89%	0	1	-6%	0.026	84%
MI-MVN	5	5	4%	0.028	89%	0	0	6%	0.034	87%	2	2	-1%	0.024	89%
MICE-PMM	0	0	5%	0.027	91%	0	0	6%	0.033	88%	25	25	12%	0.040	93%
MICE-RF	0	0	12%	0.029	94%	0	0	8%	0.032	91%	11	16	31%	0.061	98%
BSEM	0	2	7%	0.029	82%	0	6	10%	0.040	85%	0	1	1%	0.031	83%
N = 600, Mprop = 15%															
RFIML	0	0	1%	0.011	92%	0	0	1%	0.012	92%	0	0	-10%	0.010	86%
MI-MVN	0	0	1%	0.010	90%	0	0	1%	0.012	90%	0	0	-7%	0.009	88%
MICE-PMM	0	0	2%	0.011	92%	0	0	3%	0.011	91%	0	0	4%	0.013	93%
MICE-RF	0	0	4%	0.011	94%	0	0	3%	0.012	92%	0	0	19%	0.019	98%
BSEM	0	0	2%	0.011	81%	0	0	2%	0.012	81%	0	0	-8%	0.010	79%
N = 600, Mprop = 30%															
RFIML	0	0	2%	0.013	92%	0	0	4%	0.015	92%	0	0	-10%	0.012	85%
MI-MVN	0	0	2%	0.013	90%	0	0	3%	0.015	90%	0	0	-4%	0.011	88%
MICE-PMM	0	0	4%	0.013	92%	0	0	5%	0.014	91%	0	0	7%	0.021	93%
MICE-RF	0	0	9%	0.014	95%	0	0	5%	0.014	93%	0	0	25%	0.031	98%
BSEM	0	0	3%	0.014	81%	0	0	5%	0.016	82%	0	0	-7%	0.012	77%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%.

Parameter estimate bias. For mildly non-normal data, the relative biases from all methods were acceptable, regardless of the missing data mechanism, except that under MAR-Tail, MICE-RF produced substantially positive biases, especially with 30% missing data (18% - 23%). For moderately and severely non-normal data, the patterns were similar to those for mildly non-normal data, except that the biases of all methods slightly increased. MICE-RF still had the

largest biases (14% - 31%) when 30% data were missing due to MAR-Tail. Biases of the other methods were all within the acceptable range.

MSE. For all methods, MSE generally increased as the degree of non-normality increased, sample size decreased, or missing data proportion increased. The effects of missing data mechanism differed across different methods. Specifically, from MCAR to MAR-Head, MSEs generally increased for all methods; however, under MAR-Tail, the MSEs of RFIML, MI-MVN and normal-BSEM tended be the same or slightly lower than those under MACR, while the MSEs of MICE-PMM and MICE-RF became larger than those under MAR-Head. Across all conditions, MICE-RF tended to yield the largest MSEs.

CIC. For mildly non-normal data, no CIC fell below 90%. Comparing the CICs among the methods, CICs from normal-BSEM were lowest (90% - 95%) in all cells, and CICs from MICE-RF were highest (94% - 99%). Similar patterns were found for normal-BSEM and MICE-RF for moderately non-normal data. In addition, the CICs from normal-BSEM fell below 90% under MAR-Tail; and the RFIML coverages were lower than 90% when data were 30% missing due to MAR-Tail. When the population distribution was severely non-normal, more problems were observed. CICs from MI-MVN dropped below 90% when sample size was small or missing data mechanism was MAR-Tail; RFIML performed a little better than MI-MVN, but still produced coverages of less than 90% with small sample size or MAR-Tail; MICE-PMM only had problems when the sample size was small and the missingness was due to MAR-Head; CICs from MICE-RF varied largely from 90% - 98%; and finally, normal-BSEM yielded the lowest coverage across all methods (77% - 85%).

Chapter 6: Study II - Ordinal Data

Research Questions

In this study, I evaluated the performance of the methods for dealing with ordinal missing data in SEM. Based on the discussion in Chapters 2 - 4, seven methods of three classes were included in the evaluation: robust FIML (RFIML), DWLS for categorical data (cat-DWLS), normal-theory-based Bayesian SEM (BSEM), multivariate normal imputation (MI-MVN), parametric MICE (MICE- LOGIT), MICE with random forests (MICE-RF), and latent variable imputation (MI-LV). Table 6 shows a brief summary of the characteristics of the seven methods.

Table 6. Summary of Methods for Missing Ordinal Data

Method Label	Method Description	Strategy for Ordinal Data	Strategy for Missing Data	Software
RFIML	FIML with the rescaling strategy proposed by Yuan and Bentler (2000).	Rescaling (Continuous-Data)	Direct Method	R
MI-MVN	Multivariate normal imputation using EMB algorithm (Honaker, King, & Blackwell, 2011). Robust ML (RML; Satorra & Bentler, 1994) is used in the analysis phase.	Rescaling (Continuous-Data)	Imputation Method	R
cat-DWLS	Diagonally weighted least squares (Muthén & Muthén, 2008-2012).	Rescaling (Categorical-Data)	Direct Method	R
MICE-LOGIT	MICE with logistic regression model for dichotomous variables, and with cumulative logistic regression model for polytomous variables (van Buuren, 2012). Cat-DWLS (Muthén & Muthén, 2008-2012) is used in the analysis phase.	Rescaling (Categorical-Data)	Imputation Method	R
MICE-RF	MICE with random forests, which involves a random selection of a smaller group of predictors at each split (Doove et al., 2014). Cat-DWLS (Muthén & Muthén, 2008-2012) is used in the analysis phase.	Rescaling (Categorical-Data)	Imputation Method	R

MI-LV	Latent variable Imputation (Asparouhov, & Muthén, 2010c). Cat-DWLS (Muthén & Muthén, 2008-2012) is used in the analysis phase.	Rescaling (Categorical-Data)	Imputation Method	Mplus
BSEM	Gibbs sampler with data augmentation (Lee, 2007; Asparouhov, & Muthén, 2010b).	Bayesian Method	Direct Method	Mplus

The following research questions were addressed in the study.

Question 4: Are the continuous-data methods RFIML and MI-MVN applicable to ordinal data? Under what situations and to what extent are the two methods robust to discontinuity?

Question 5: Do normal-theory-based BSEM and MI-LV perform well under a broader range of conditions than those examined in Asparouhov and Muthén (2010a)?

Question 6: How are the methods influenced by sample size, degree of non-normality, missing data mechanism and missing data proportion?

Question 7: Which of the seven methods performs best under the examined conditions?

Method

Data generation model. The data generation model in Study I was also used in this study. The correlation matrix derived from given parameter values was used for generating ordinal data.

Design factors.

Number of categories. Both dichotomous data and polytomous ordinal data were included in the study. The numbers of ordinal categories were set at 2, 3, 5, and 7.

Asymmetry of thresholds. Ordinal data were generated using the method proposed by Ferrari and Barbiero (2012). This method first generates multivariate normal data based on the population correlation matrix and then discretizes them according to the user-specified cumulative probability for each variable. Thus, the asymmetry of the thresholds and the number

of categories can be easily specified. For the sake of simplicity, all variables had the same degree of asymmetry and the same number of categories. Three degrees of asymmetry (symmetry, moderate asymmetry and severe asymmetry) were specified following Rhemtulla, Brosseau-Liard, and Savalei (2012) and Wu et al. (2015).

Sample size. Two levels of sample size were examined: small (300) and large (600).

Missing data proportion. Missing data proportions were manipulated at two levels: small (15%) and high (30%)

Missing data mechanism. Same as Study I, three missing data mechanisms were created in this study: MCAR, MAR-Head and MAR-Tail.

One thousand replicated samples were created in each of the fully-crossed conditions ($4 \times 3 \times 2 \times 2 \times 3 = 144$). In addition, in order to differentiate the impacts of discontinuity and missingness, I evaluated the three direct methods (i.e., RML, cat-DWLS, and normal-BSEM) under the conditions with complete data and all levels of number of categories, asymmetry of thresholds and sample size ($4 \times 3 \times 2 = 24$). The analysis model was the same as the data generation model. For the imputation methods, 50 imputed data sets were obtained for each replication following the guideline of White et al. (2011). Figures 3 - 6 demonstrate the distributions of one manifest variable for one replication with various numbers of categories, degrees of asymmetry, and missing data mechanisms.

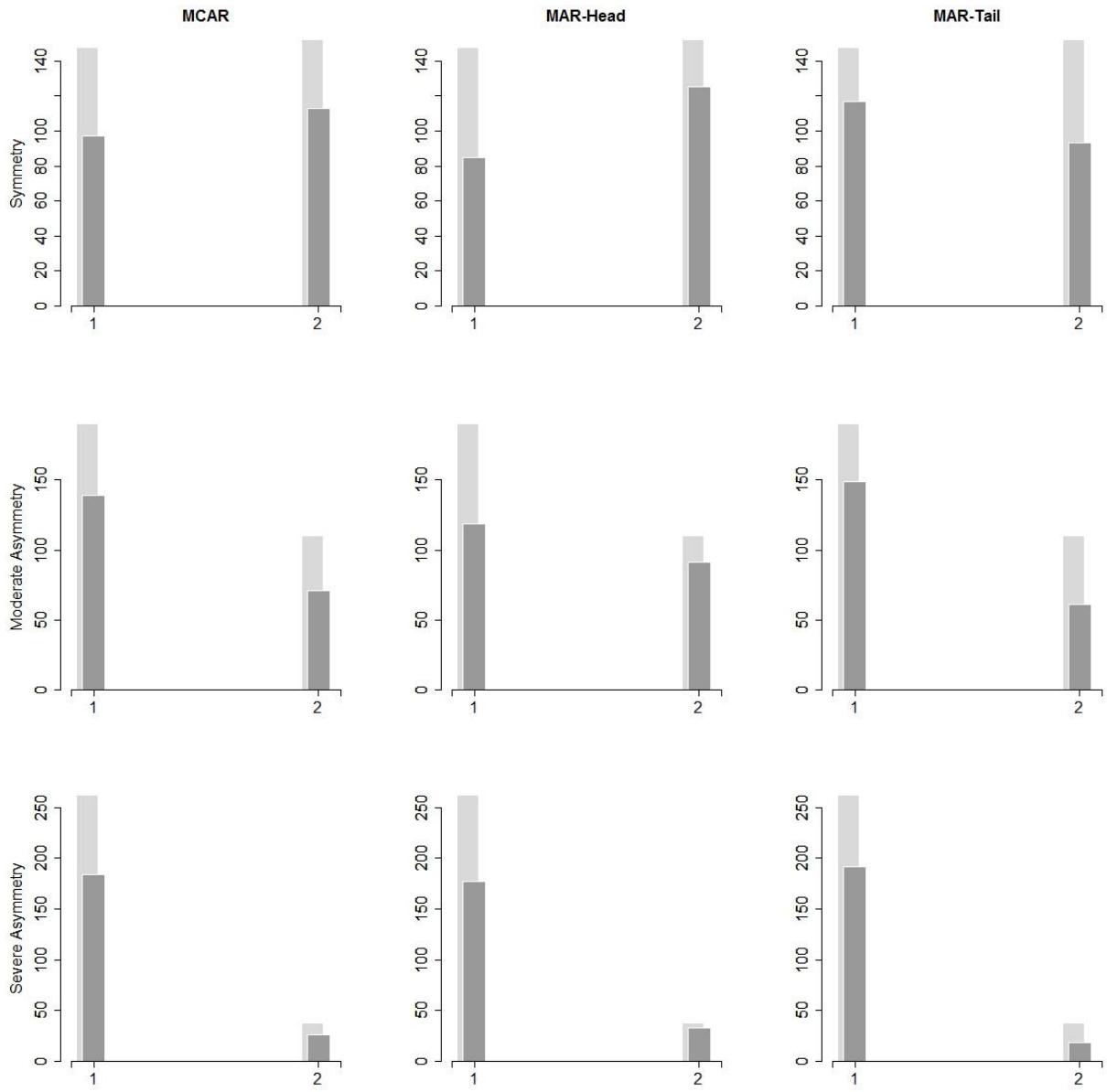

Figure 3. Distributions of x1 (two categories) for one replication with N = 300 before (light grey) and after (dark grey) imposing 30% missing data.

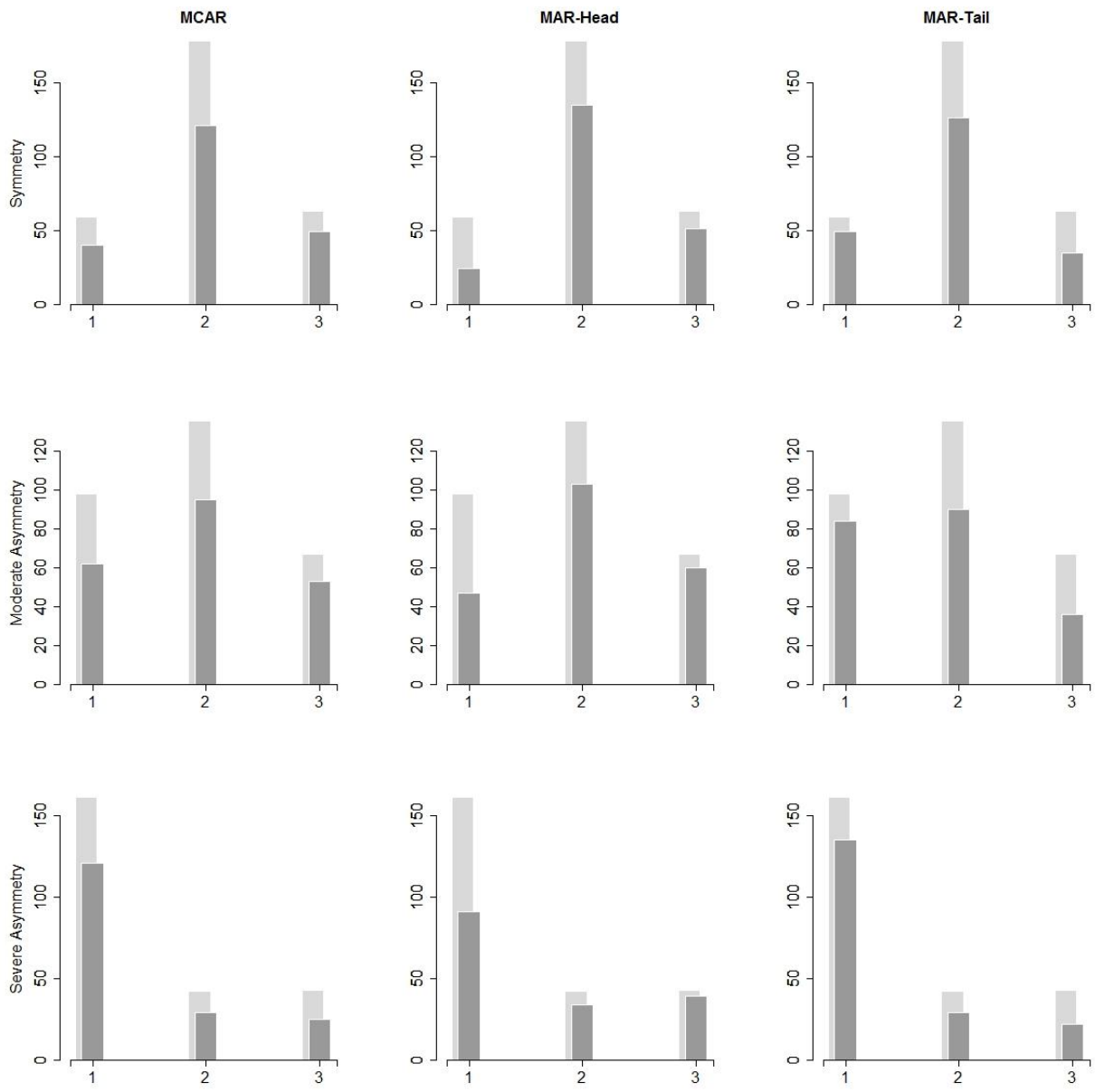

Figure 4. Distributions of x1 (three categories) for one replication with N = 300 before (light grey) and after (dark grey) imposing 30% missing data.

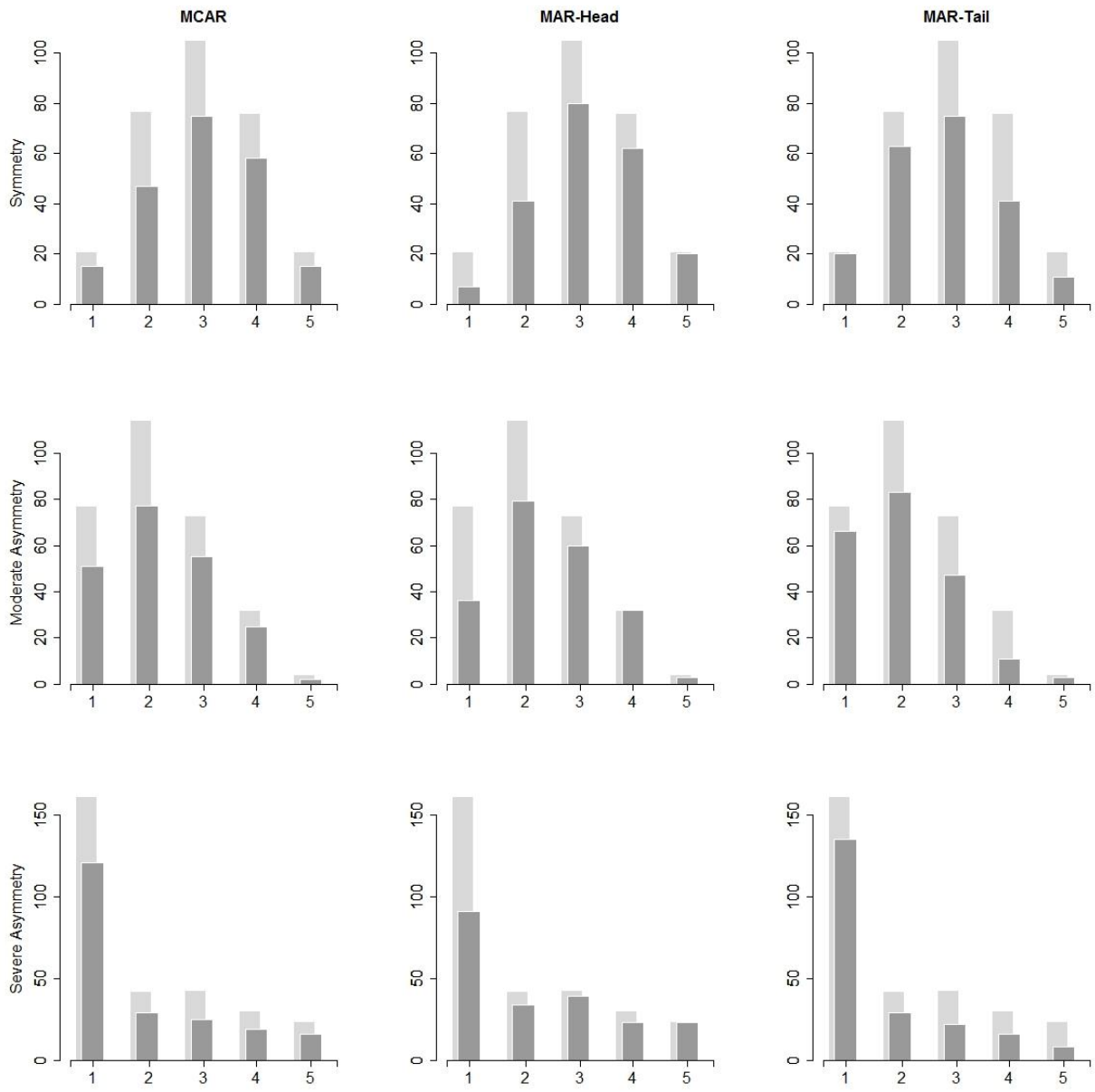

Figure 5. Distributions of x1 (five categories) for one replication with N = 300 before (light grey) and after (dark grey) imposing 30% missing data.

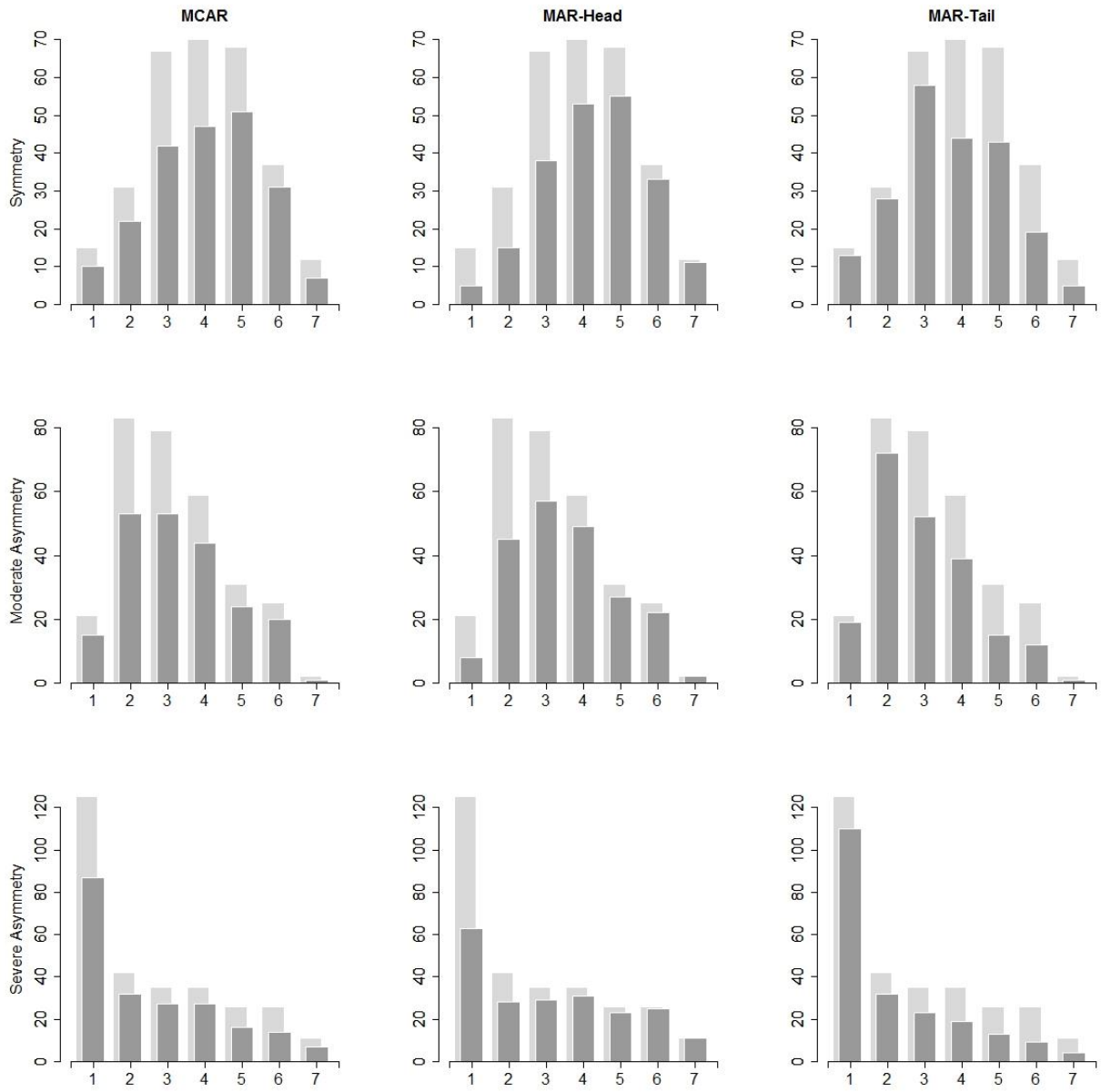

Figure 6. Distributions of x1 (seven categories) for one replication with N = 300 before (light grey) and after (dark grey) imposing 30% missing data.

Computational characteristics. Data were generated in R (R Core Team, 2014) using functions provided in package GenOrd (Ferrari & Barbiero, 2014). MI-LV was implemented through Mplus 7.0 (Muthén & Muthén, 2008-2012). The computational characteristics for robust FIML, normal-BSEM, MI-MVN, and the MICE methods in this study were the same as those described in Study I, except that 1) the prior for any of the loadings in normal-BSEM was $N(0,5)$ by default in Mplus; and 2) the imputed data sets were analyzed using cat-DWLS in lavaan (Rosseel, 2012).

Results

Findings are summarized with respect to the nature of discontinuity. Dichotomous data results were described first and followed by those for the polytomous data. Consistent with Study I, the performance of the seven missing ordinal data methods was evaluated based on four outcomes: convergence failures and outliers, relative parameter estimate bias, mean squared Error (MSE) and confidence interval (or credible interval) coverage (CIC).

The results of the three direct complete data methods were first summarized for the purpose of comparison (Table 7). Almost 100% of the replications successfully converged under each design cell for all three methods and there were very few outliers. RML performed the best (smallest biases and MSEs and appropriate CICs) across all cells on all four outcomes; cat-DWLS tended to yield large bias when thresholds were severely asymmetrical and the number of categories were small (2 or 3); normal-BSEM produced substantially large bias and MSE with small numbers of categories (2 or 3) or small sample size (N = 300), and was less affected by the asymmetry of thresholds. The results for missing ordinal data are described as follows.

Table 7. Results for Complete Ordinal Data

	Symmetry					Moderate Asymmetry					Severe Asymmetry				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
							Ncat = 2, N = 300								
RML	0	0	2%	0.015	96%	0	0	2%	0.017	95%	3	3	3%	0.038	94%
cat-DWLS	0	0	7%	0.016	96%	0	0	7%	0.018	95%	3	3	18%	0.040	95%
BSEM	0	0	78%	0.429	97%	0	0	64%	0.267	98%	0	5	69%	0.191	99%
							Ncat = 2, N = 600								
RML	0	0	1%	0.007	95%	0	0	1%	0.008	95%	0	0	3%	0.016	94%
cat-DWLS	0	0	5%	0.008	95%	0	0	6%	0.009	96%	0	0	17%	0.019	94%
BSEM	0	3	13%	0.018	96%	0	2	17%	0.023	97%	0	0	41%	0.066	98%
							Ncat = 3, N = 300								
RML	0	0	1%	0.011	95%	0	0	1%	0.010	95%	0	0	1%	0.017	94%
cat-DWLS	0	0	2%	0.011	95%	0	0	3%	0.010	95%	0	0	11%	0.019	95%
BSEM	0	1	22%	0.089	96%	0	0	19%	0.067	96%	0	0	89%	0.434	97%
							Ncat = 3, N = 600								
RML	0	0	1%	0.006	95%	0	0	1%	0.005	95%	0	0	1%	0.009	95%
cat-DWLS	0	0	2%	0.005	94%	0	0	2%	0.005	94%	0	0	10%	0.009	95%
BSEM	0	2	6%	0.010	95%	0	0	5%	0.009	96%	0	3	20%	0.023	96%
							Ncat = 5, N = 300								
RML	0	0	1%	0.009	95%	0	0	1%	0.009	95%	0	0	0%	0.011	95%
cat-DWLS	0	0	2%	0.008	95%	0	0	2%	0.009	96%	0	0	6%	0.011	95%
BSEM	0	3	6%	0.015	96%	0	0	8%	0.022	96%	0	0	14%	0.032	96%
							Ncat = 5, N = 600								
RML	0	0	1%	0.004	95%	0	0	1%	0.005	95%	0	0	1%	0.006	96%
cat-DWLS	0	0	1%	0.004	95%	0	0	2%	0.004	95%	0	0	6%	0.006	95%
BSEM	0	0	2%	0.006	95%	0	0	3%	0.007	96%	0	0	8%	0.009	95%
							Ncat = 7, N = 300								
RML	0	0	1%	0.008	95%	0	0	1%	0.008	96%	0	0	1%	0.010	95%
cat-DWLS	0	0	2%	0.008	96%	0	0	2%	0.008	97%	0	0	5%	0.009	96%
BSEM	0	2	5%	0.013	95%	0	0	5%	0.014	95%	0	0	9%	0.019	96%
							Ncat = 7, N = 600								
RML	0	0	1%	0.004	95%	0	0	1%	0.004	95%	0	0	1%	0.005	95%
cat-DWLS	0	0	1%	0.004	95%	0	0	2%	0.004	96%	0	0	4%	0.005	95%
BSEM	0	0	3%	0.006	95%	0	0	3%	0.006	95%	0	0	6%	0.008	95%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RML's MSE under the same condition, or CIC < 90%.

Dichotomous data.

Convergence failures and outliers. For missing dichotomous data, the convergence rate was highly impacted by asymmetry of threshold, missing data proportion and sample size. If the convergence rate fell below 90%, the method was considered failed in that condition, and

therefore the other three outcomes (bias, MSE and CIC) were not shown in the tables (Tables 8 - 10). Under the symmetric threshold, the convergence rates of all seven methods were higher than 94% (see Table 8), regardless of missing data proportion, sample size and missing data mechanism. However, for data with a moderately asymmetrical threshold and 30% missingness, the greatest numbers of convergence failures were found associated with MICE-RF (705 and 951); substantive convergence failures were also observed for the other R-implemented methods (i.e., RFIML, MI-MVN, cat-DWLS and MICE-LOGIT), especially with small sample size (281 - 769 at N = 300, and 60 - 207 at N = 600, see Table 9). The lowest convergence rates were observed under the severely asymmetrical threshold. More specifically, all the methods implemented in R had only 0% - 7% convergence rate (Table 10) when the missing data occurred due to MAR-Tail; the R implemented imputation methods (i.e., MI-MVN, MICE-LOGIT, MICE-RF) also encountered convergence problems with the small sample size even under MCAR or MAR-Head. In comparison, the Mplus- implemented methods (i.e., MI-LV and normal-BSEM) had a 100% convergence rate. The numbers of outliers were trivial across all conditions.

Parameter estimate bias. The patterns in bias were similar under the conditions of symmetric threshold and moderately asymmetrical threshold (Table 8 -Table 9), in which RFIML and MI-MVN yielded negligible biases across all conditions (< 5%) ; normal-BSEM produced the highest biases (16% - 213%) across all methods in all design cells; cat-DWLS and MICE-RF were sensitive to missing data proportion and missing data mechanism. Under severely asymmetrical threshold (Table 10) and MAR-Tail, biases of the R-implemented methods could not be computed due to the extremely low convergence rates; the Mplus-implemented methods also produced large biases ranging from 36% to 169%. Under severely

asymmetrical threshold and MCAR or MAR-Head, RFIML still yielded the lowest biases (4% - 8%); and MI-MVN's performance was as good as RFIML. However, the parameter estimates of all the other five methods were found to be biased. The only exception was MICE-RF, which only produced biased estimates under MCAR, but not MAR-Head.

Table 8. Results for Missing Dichotomous Data with Symmetric Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	0	0	3%	0.019	95%	0	0	2%	0.018	95%	0	0	3%	0.018	96%
MI-MVN	0	0	3%	0.019	94%	0	0	3%	0.018	95%	0	0	3%	0.019	94%
cat-DWLS	0	0	8%	0.020	95%	0	0	11%	0.021	96%	0	0	11%	0.022	96%
MICE-LOGIT	1	1	8%	0.020	95%	0	0	8%	0.019	95%	0	0	8%	0.020	95%
MICE-RF	0	0	9%	0.021	97%	0	0	9%	0.020	97%	0	0	10%	0.021	97%
MI-LV	0	0	8%	0.020	96%	0	0	8%	0.019	96%	0	0	8%	0.020	96%
BSEM	0	0	104%	0.681	97%	0	0	90%	0.549	97%	0	0	95%	0.605	98%
N = 300, Mprop = 30%															
RFIML	0	0	3%	0.024	95%	0	0	4%	0.025	95%	0	1	4%	0.026	95%
MI-MVN	41	42	3%	0.027	95%	37	38	5%	0.028	94%	37	38	5%	0.029	95%
cat-DWLS	0	0	9%	0.026	95%	1	2	24%	0.046	96%	0	1	23%	0.048	96%
MICE-LOGIT	9	10	9%	0.027	95%	7	8	10%	0.027	94%	3	3	10%	0.029	94%
MICE-RF	41	42	12%	0.029	99%	57	65	20%	0.042	99%	40	54	20%	0.045	99%
MI-LV	0	0	10%	0.028	96%	0	0	12%	0.029	96%	0	0	12%	0.032	96%
BSEM	0	3	150%	1.036	97%	0	3	146%	0.921	97%	0	6	149%	1.001	97%
N = 600, Mprop = 15%															
RFIML	0	0	1%	0.009	94%	0	0	2%	0.009	95%	0	0	1%	0.009	95%
MI-MVN	0	0	1%	0.009	94%	0	0	2%	0.009	95%	0	0	2%	0.009	94%
cat-DWLS	0	0	5%	0.010	94%	0	0	9%	0.011	95%	0	0	9%	0.011	94%
MICE-LOGIT	0	0	5%	0.010	94%	0	0	6%	0.010	95%	0	0	6%	0.010	94%
MICE-RF	0	0	6%	0.010	96%	0	0	7%	0.010	96%	0	0	6%	0.010	96%
MI-LV	0	0	6%	0.010	94%	0	0	6%	0.010	95%	0	0	6%	0.010	94%
BSEM	0	2	20%	0.046	96%	0	2	16%	0.027	96%	0	5	16%	0.031	96%
N = 600, Mprop = 30%															
RFIML	0	0	1%	0.011	95%	0	0	2%	0.011	96%	0	0	2%	0.012	95%
MI-MVN	0	0	2%	0.012	94%	0	0	3%	0.012	95%	0	0	3%	0.013	94%
cat-DWLS	0	0	6%	0.012	95%	0	0	20%	0.022	94%	0	0	20%	0.023	94%
MICE-LOGIT	0	0	6%	0.013	94%	0	0	7%	0.012	94%	0	0	7%	0.013	93%
MICE-RF	0	0	8%	0.013	98%	0	0	14%	0.019	98%	0	0	15%	0.020	98%
MI-LV	0	0	7%	0.013	95%	0	0	7%	0.013	96%	0	0	7%	0.013	95%
BSEM	0	2	28%	0.082	96%	0	1	25%	0.054	97%	0	1	30%	0.074	96%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%.

Table 9. Results for Missing Dichotomous Data with Moderately Asymmetrical Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	0	0	2%	0.021	95%	0	0	2%	0.020	95%	0	0	2%	0.023	95%
MI-MVN	0	0	2%	0.022	94%	0	0	2%	0.021	94%	3	3	3%	0.024	93%
cat-DWLS	0	0	8%	0.023	95%	0	0	11%	0.023	95%	0	0	14%	0.029	95%
MICE-LOGIT	0	0	9%	0.023	95%	0	0	8%	0.022	95%	0	1	9%	0.024	95%
MICE-RF	1	1	9%	0.024	97%	0	0	5%	0.019	96%	13	15	18%	0.038	98%
MI-LV	0	0	9%	0.024	95%	0	0	8%	0.022	95%	0	0	10%	0.026	96%
BSEM	0	1	83%	0.402	97%	0	0	71%	0.349	97%	0	0	92%	0.477	98%
N = 300, Mprop = 30%															
RFIML	0	0	3%	0.026	95%	0	0	3%	0.025	96%	281	-	-	-	-
MI-MVN	63	67	3%	0.029	94%	10	10	4%	0.028	95%	769	-	-	-	-
cat-DWLS	0	0	10%	0.029	95%	0	0	18%	0.037	95%	282	-	-	-	-
MICE-LOGIT	20	24	10%	0.030	94%	1	2	9%	0.029	95%	327	-	-	-	-
MICE-RF	81	83	12%	0.032	98%	1	1	6%	0.024	98%	951	-	-	-	-
MI-LV	0	1	12%	0.033	96%	0	0	10%	0.030	96%	0	14	18%	0.057	97%
BSEM	0	2	123%	0.697	98%	0	1	100%	0.563	97%	0	56	213%	1.418	99%
N = 600, Mprop = 15%															
RFIML	0	0	1%	0.009	95%	0	0	2%	0.009	95%	0	0	2%	0.010	95%
MI-MVN	1	1	1%	0.009	94%	0	0	2%	0.009	94%	0	0	2%	0.010	94%
cat-DWLS	0	0	6%	0.010	95%	0	0	10%	0.011	95%	0	0	12%	0.014	95%
MICE-LOGIT	0	0	7%	0.010	95%	0	0	7%	0.010	95%	0	0	8%	0.012	95%
MICE-RF	1	1	7%	0.011	97%	0	0	3%	0.009	97%	0	0	16%	0.018	97%
MI-LV	0	0	7%	0.010	96%	0	0	7%	0.010	95%	0	0	8%	0.012	95%
BSEM	0	3	22%	0.043	97%	0	2	18%	0.030	97%	0	0	27%	0.071	97%
N = 600, Mprop = 30%															
RFIML	0	0	2%	0.012	95%	0	0	2%	0.011	95%	60	60	4%	0.018	95%
MI-MVN	1	1	2%	0.013	94%	0	0	3%	0.012	94%	207	-	-	-	-
cat-DWLS	0	0	8%	0.013	95%	0	0	17%	0.018	94%	60	63	39%	0.071	95%
MICE-LOGIT	1	1	8%	0.014	94%	0	0	7%	0.013	94%	64	64	10%	0.018	92%
MICE-RF	5	5	9%	0.014	98%	0	0	3%	0.011	98%	705	-	-	-	-
MI-LV	0	0	9%	0.014	95%	0	0	8%	0.013	95%	0	0	12%	0.020	97%
BSEM	0	0	32%	0.070	97%	0	1	23%	0.046	97%	0	3	64%	0.200	98%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%. Bias, MSE and CIC are not computed if Convergence Failures ≥ 100.

Table 10. Results for Missing Dichotomous Data with Severely Asymmetrical Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	12	12	5%	0.055	93%	1	2	4%	0.045	94%	1000	-	-	-	-
MI-MVN	113	-	-	-	-	37	38	6%	0.046	91%	1000	-	-	-	-
cat-DWLS	6	6	20%	0.055	95%	1	1	21%	0.049	95%	1000	-	-	-	-
MICE-LOGIT	49	55	22%	0.053	95%	7	8	20%	0.049	95%	1000	-	-	-	-
MICE-RF	60	64	20%	0.055	97%	8	8	13%	0.039	97%	1000	-	-	-	-
MI-LV	0	2	23%	0.063	96%	0	1	21%	0.050	95%	0	195	36%	0.171	98%
BSEM	0	16	76%	0.253	99%	0	6	65%	0.192	99%	0	251	132%	0.695	99%
N = 300, Mprop = 30%															
RFIML	55	57	7%	0.077	93%	7	7	8%	0.059	94%	998	-	-	-	-
MI-MVN	619	-	-	-	-	205	-	-	-	-	998	-	-	-	-
cat-DWLS	55	58	22%	0.069	94%	1	2	27%	0.067	95%	998	-	-	-	-
MICE-LOGIT	303	-	-	-	-	52	61	24%	0.063	96%	998	-	-	-	-
MICE-RF	488	-	-	-	-	43	45	9%	0.037	98%	998	-	-	-	-
MI-LV	0	23	30%	0.095	97%	0	2	26%	0.069	97%	0	234	36%	0.209	99%
BSEM	0	37	101%	0.418	99%	0	20	80%	0.266	99%	0	380	169%	1.050	100%
N = 600, Mprop = 15%															
RFIML	0	0	3%	0.021	94%	0	0	4%	0.019	94%	931	-	-	-	-
MI-MVN	2	2	3%	0.021	91%	0	0	5%	0.020	93%	998	-	-	-	-
cat-DWLS	0	0	17%	0.024	94%	0	0	19%	0.024	94%	931	-	-	-	-
MICE-LOGIT	0	0	18%	0.024	94%	0	0	17%	0.022	95%	940	-	-	-	-
MICE-RF	2	2	17%	0.024	98%	0	0	11%	0.018	97%	1000	-	-	-	-
MI-LV	0	0	18%	0.024	95%	0	0	18%	0.023	95%	0	62	38%	0.108	98%
BSEM	0	0	47%	0.093	98%	0	1	39%	0.067	98%	0	71	125%	0.488	98%
N = 600, Mprop = 30%															
RFIML	0	0	4%	0.024	94%	0	0	6%	0.023	95%	1000	-	-	-	-
MI-MVN	46	46	4%	0.026	92%	4	4	7%	0.025	93%	1000	-	-	-	-
cat-DWLS	0	0	18%	0.029	95%	0	0	24%	0.031	94%	1000	-	-	-	-
MICE-LOGIT	18	18	19%	0.030	94%	0	1	18%	0.026	95%	1000	-	-	-	-
MICE-RF	69	69	18%	0.031	99%	1	1	5%	0.017	98%	1000	-	-	-	-
MI-LV	0	0	21%	0.032	96%	0	0	20%	0.027	95%	0	97	39%	0.109	99%
BSEM	0	1	53%	0.109	98%	0	1	42%	0.074	98%	0	54	125%	0.621	99%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%. Bias, MSE and CIC are not computed if Convergence Failures ≥ 100.

MSE and CIC. For all evaluated methods, MSE generally decreased as the sample size increased, the missing data proportion decreased and the threshold became more asymmetrical. Across all design cells, normal-BSEM yielded the highest MSEs among the seven methods. The CICs were above 90% across all conditions, except for those conditions in which CICs were not computed due to large numbers of convergence failures. The lowest CICs were from MI-MVN under MAR-Tail, which ranged from 91% to 94%, and the highest CICs were from MICE-RF and normal-BSEM under all conditions, which ranged from 96% to 99%. CICs from other methods were all equal or close to 95%.

Polytomous data.

Convergence failures and outliers. For missing polytomous data, the R-implemented methods still had convergence problems (convergence rate < 90%), however, a variety of patterns were observed for different methods (Table 11 - Table 19). RFIML and cat-DWLS had convergence problems only for three-category data with severely asymmetrical thresholds and 30% MAR-Tail missingness; MI-MVN failed to converge with the small sample size and the large missing data proportion, however, the problems were minimized as the number of ordinal categories increased; MICE-LOGIT performed well only with the large sample size, and the small proportion of missing data, particularly under MAR-Head, as the number of categories increased. The most severe convergence problems were found with MICE-RF, which failed to converge with 30% missingness under MAR-Tail, regardless of sample size and number of categories. All the Mplus-implemented methods (i.e., MI-LV and normal-BSEM) converged at a rate of 100%. The numbers of outliers were trivial across all conditions.

Parameter estimate bias. Under the conditions of symmetric thresholds and moderately asymmetrical thresholds, only MICE-RF and normal-BSEM produced unacceptable biases

(Table 11 - Table 19). Specifically, MICE-RF was sensitive to the amount of data and missing data mechanism, and biases were mostly found under MAR-Head and MAR-Tail with 30% missingness, or with 15% missingness when the sample size was small. On the other hand, normal-BSEM seemed to be only sensitive to y sample size, because biases were mostly observed at $N = 300$. Under severely asymmetrical threshold, different patterns were observed for different numbers of categories. For data with three categories (Table 11 - Table 13), RFIML and MI-MVN performed well (bias $\leq 5\%$) across sample size, missing data proportion, and missing data mechanism (except for cells with convergence problems); normal-BSEM produced the highest biases (20% - 228%); cat-DWLS became problematic under MAR-Head and MAR-Tail (either non-converged or bias exceeded 15%); MICE-RF surprisingly had a quite small bias (2% -6%) under MAR-Head; the biases for other methods or other conditions exceeded 10%, but still within an acceptable range, according to Muthén, et al. (1987). For data with five or seven categories (Table 14 - Table 19), only MICE-RF and normal-BSEM were found problematic under a limited number of conditions. For example, for MAR-Tail data, MICE-RF yielded either large bias or large numbers of convergence failures; biases yielded by normal-BSEM had been largely reduced compared to those with three-category data, however, normal-BSEM still did not work well at $N = 300$.

MSE and CIC. The result of MSE and CIC for polytomous data followed a similar pattern as the dichotomous data. In addition, MSEs decreased as the number of categories increased (Table 11 - Table 19). The MSEs from all methods were comparable across design cells, except for normal-BSEM and MICE-RF. Compared to the other methods, normal-BSEM yielded substantially larger MSEs unless $N = 600$ and there were 15% data missing. MICE-RF tended to produce larger MSEs than the other five methods under MAR-Head and MAR-Tail

with symmetric and moderately asymmetrical thresholds, and MAR-Tail data with severely asymmetrical thresholds. CICs from the examined methods were not influenced much by the design factors. MICE-RF yielded the highest CICs (96% - 100%) across all design cells, while MICE-LOGIT tended to have the lowest CICs (88% - 94%) under MAR-Head and MAR-Tail. CICs for other methods were all close to 95%.

Table 11. Results for Missing Three-Category Data with Symmetric Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	0	0	1%	0.013	95%	0	0	2%	0.015	95%	0	0	3%	0.016	95%
MI-MVN	0	0	1%	0.014	94%	1	1	2%	0.015	94%	2	4	3%	0.016	94%
cat-DWLS	0	0	3%	0.013	95%	0	0	3%	0.015	94%	0	0	3%	0.016	95%
MICE-LOGIT	0	0	3%	0.013	95%	10	11	4%	0.016	93%	4	4	4%	0.017	94%
MICE-RF	0	0	3%	0.014	97%	7	8	14%	0.026	98%	5	6	15%	0.029	98%
MI-LV	0	0	3%	0.013	95%	0	0	3%	0.014	95%	0	0	4%	0.015	95%
BSEM	0	1	35%	0.122	97%	0	0	49%	0.209	96%	0	0	49%	0.194	97%
N = 300, Mprop = 30%															
RFIML	0	0	1%	0.016	95%	0	0	2%	0.019	95%	0	0	3%	0.020	96%
MI-MVN	3	3	1%	0.017	94%	9	9	3%	0.020	95%	11	11	4%	0.021	95%
cat-DWLS	0	0	3%	0.016	94%	0	0	8%	0.023	96%	0	0	10%	0.025	96%
MICE-LOGIT	1	1	4%	0.017	95%	67	70	5%	0.023	92%	80	83	5%	0.022	93%
MICE-RF	12	12	5%	0.018	99%	22	28	20%	0.040	99%	49	58	22%	0.046	100%
MI-LV	0	0	4%	0.017	95%	0	0	5%	0.018	95%	0	0	6%	0.019	96%
BSEM	0	0	75%	0.451	95%	0	0	90%	0.600	95%	0	2	95%	0.651	95%
N = 600, Mprop = 15%															
RFIML	0	0	1%	0.007	94%	0	0	3%	0.020	96%	0	0	2%	0.007	95%
MI-MVN	0	0	1%	0.007	94%	0	0	2%	0.008	94%	0	0	2%	0.008	95%
cat-DWLS	0	0	2%	0.007	94%	0	0	2%	0.008	95%	0	0	2%	0.007	95%
MICE-LOGIT	0	0	2%	0.007	94%	0	0	3%	0.008	93%	0	0	3%	0.008	94%
MICE-RF	0	0	2%	0.007	97%	0	0	12%	0.013	97%	0	0	12%	0.013	97%
MI-LV	0	0	2%	0.007	94%	0	0	2%	0.007	94%	0	0	3%	0.007	95%
BSEM	0	2	5%	0.012	95%	0	3	7%	0.015	95%	0	4	8%	0.015	96%
N = 600, Mprop = 30%															
RFIML	0	0	2%	0.008	95%	0	0	2%	0.009	95%	0	0	3%	0.009	96%
MI-MVN	0	0	2%	0.008	94%	0	0	3%	0.010	95%	0	0	3%	0.010	95%
cat-DWLS	0	0	2%	0.008	95%	0	0	7%	0.011	96%	0	0	8%	0.011	96%
MICE-LOGIT	0	0	3%	0.008	94%	3	3	3%	0.011	91%	2	2	4%	0.011	92%
MICE-RF	0	0	3%	0.008	98%	0	0	17%	0.019	99%	0	0	17%	0.020	99%
MI-LV	0	0	3%	0.008	94%	0	0	3%	0.009	95%	0	0	3%	0.008	95%
BSEM	0	3	8%	0.019	95%	0	3	11%	0.024	95%	0	1	13%	0.029	95%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%.

Table 12. Results for Missing Three-Category Data with Moderately Asymmetrical Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	0	0	1%	0.012	95%	0	0	1%	0.012	95%	0	0	2%	0.014	95%
MI-MVN	0	0	1%	0.012	94%	0	0	1%	0.013	95%	0	0	2%	0.014	95%
cat-DWLS	0	0	4%	0.013	95%	0	0	5%	0.013	95%	0	0	4%	0.015	95%
MICE-LOGIT	0	0	4%	0.012	95%	0	0	3%	0.013	95%	0	0	3%	0.014	94%
MICE-RF	0	0	4%	0.013	97%	0	0	8%	0.016	97%	1	1	12%	0.022	98%
MI-LV	0	0	4%	0.012	95%	0	0	3%	0.012	95%	0	0	4%	0.013	95%
BSEM	0	0	33%	0.142	97%	0	2	32%	0.127	96%	0	0	41%	0.174	97%
N = 300, Mprop = 30%															
RFIML	0	0	1%	0.015	95%	0	0	2%	0.017	95%	0	0	4%	0.018	96%
MI-MVN	1	1	1%	0.016	94%	10	10	3%	0.018	95%	8	8	4%	0.020	96%
cat-DWLS	0	0	4%	0.016	94%	0	0	12%	0.025	96%	0	0	13%	0.027	96%
MICE-LOGIT	0	0	4%	0.016	94%	10	11	5%	0.020	93%	16	19	3%	0.019	92%
MICE-RF	4	4	6%	0.017	98%	31	36	21%	0.037	99%	92	105	26%	0.050	99%
MI-LV	0	0	4%	0.016	95%	0	0	4%	0.017	95%	0	0	7%	0.018	96%
BSEM	0	0	60%	0.396	95%	0	0	83%	0.594	96%	0	0	95%	0.651	96%
N = 600, Mprop = 15%															
RFIML	0	0	1%	0.006	95%	0	0	1%	0.006	95%	0	0	2%	0.007	95%
MI-MVN	0	0	1%	0.006	94%	0	0	1%	0.006	95%	0	0	2%	0.007	95%
cat-DWLS	0	0	2%	0.006	94%	0	0	4%	0.007	95%	0	0	3%	0.007	95%
MICE-LOGIT	0	0	2%	0.006	94%	0	0	2%	0.006	95%	0	0	3%	0.007	94%
MICE-RF	0	0	3%	0.006	97%	0	0	7%	0.008	97%	0	0	10%	0.010	97%
MI-LV	0	0	2%	0.006	95%	0	0	2%	0.006	95%	0	0	3%	0.007	95%
BSEM	0	1	5%	0.010	95%	0	0	6%	0.011	96%	0	3	7%	0.012	96%
N = 600, Mprop = 30%															
RFIML	0	0	1%	0.007	95%	0	0	2%	0.008	96%	0	0	2%	0.008	95%
MI-MVN	0	0	1%	0.008	95%	0	0	2%	0.008	96%	0	0	2%	0.009	94%
cat-DWLS	0	0	3%	0.008	95%	0	0	10%	0.013	96%	0	0	10%	0.013	95%
MICE-LOGIT	0	0	3%	0.008	94%	0	0	3%	0.009	93%	0	0	3%	0.009	91%
MICE-RF	0	0	4%	0.008	98%	0	0	17%	0.018	99%	0	0	19%	0.021	99%
MI-LV	0	0	3%	0.008	95%	0	0	3%	0.008	96%	0	0	4%	0.008	95%
BSEM	0	3	7%	0.015	96%	0	5	9%	0.018	96%	0	4	12%	0.026	96%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%.

Table 13. Results for Missing Three-Category Data with Severely Asymmetrical Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	0	0	1%	0.021	95%	0	0	2%	0.020	95%	6	7	3%	0.034	92%
MI-MVN	1	1	1%	0.022	93%	0	0	2%	0.021	93%	526	-	-	-	-
cat-DWLS	0	0	12%	0.023	95%	0	0	14%	0.023	94%	4	6	22%	0.045	97%
MICE-LOGIT	0	0	12%	0.023	95%	0	0	12%	0.023	95%	33	34	-3%	0.021	88%
MICE-RF	1	1	12%	0.024	97%	0	0	6%	0.018	96%	905	-	-	-	-
MI-LV	0	0	13%	0.023	95%	0	0	12%	0.022	95%	0	2	13%	0.030	97%
BSEM	0	0	113%	0.590	98%	0	0	83%	0.369	98%	0	9	170%	1.015	98%
N = 300, Mprop = 30%															
RFIML	1	1	2%	0.027	94%	0	0	3%	0.025	95%	743	-	-	-	-
MI-MVN	85	86	2%	0.030	93%	8	8	5%	0.028	94%	987	-	-	-	-
cat-DWLS	0	0	13%	0.030	95%	0	0	20%	0.033	95%	748	-	-	-	-
MICE-LOGIT	22	24	14%	0.031	95%	2	2	16%	0.030	95%	734	-	-	-	-
MICE-RF	72	73	14%	0.033	98%	0	0	4%	0.018	97%	996	-	-	-	-
MI-LV	0	1	15%	0.033	96%	0	0	14%	0.028	95%	0	34	30%	0.119	97%
BSEM	0	2	133%	0.833	97%	0	0	106%	0.661	96%	0	59	228%	1.832	96%
N = 600, Mprop = 15%															
RFIML	0	0	1%	0.010	95%	0	0	2%	0.010	95%	0	0	1%	0.015	93%
MI-MVN	0	0	1%	0.010	93%	0	0	2%	0.010	94%	64	64	1%	0.015	92%
cat-DWLS	0	0	10%	0.011	95%	0	0	13%	0.012	94%	0	0	19%	0.022	97%
MICE-LOGIT	0	0	10%	0.011	95%	0	0	10%	0.011	94%	5	6	1%	0.013	86%
MICE-RF	0	0	10%	0.012	97%	0	0	6%	0.009	96%	519	-	-	-	-
MI-LV	0	0	11%	0.011	95%	0	0	11%	0.011	95%	0	0	10%	0.014	96%
BSEM	0	2	21%	0.036	97%	0	2	20%	0.028	98%	0	1	45%	0.125	97%
N = 600, Mprop = 30%															
RFIML	0	0	1%	0.012	95%	0	0	4%	0.012	95%	392	-	-	-	-
MI-MVN	0	0	2%	0.013	93%	0	0	4%	0.013	94%	863	-	-	-	-
cat-DWLS	0	0	11%	0.014	95%	0	0	19%	0.017	94%	391	-	-	-	-
MICE-LOGIT	0	0	11%	0.014	94%	0	0	11%	0.013	95%	385	-	-	-	-
MICE-RF	1	1	11%	0.015	98%	0	0	2%	0.009	97%	999	-	-	-	-
MI-LV	0	0	12%	0.014	95%	0	0	11%	0.013	95%	0	1	19%	0.038	97%
BSEM	0	2	33%	0.079	97%	0	4	23%	0.039	97%	0	2	122%	0.599	94%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%. Bias, MSE and CIC are not computed if Convergence Failures ≥ 100.

Table 14. Results for Missing Five-Category Data with Symmetric Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	0	0	1%	0.010	95%	0	0	1%	0.011	95%	0	0	2%	0.011	95%
MI-MVN	0	0	1%	0.010	95%	0	0	1%	0.011	95%	0	0	2%	0.012	94%
cat-DWLS	0	0	2%	0.010	95%	0	0	0%	0.011	95%	0	0	1%	0.011	96%
MICE-LOGIT	0	0	2%	0.010	95%	2	2	5%	0.014	94%	0	0	7%	0.021	91%
MICE-RF	0	0	3%	0.010	97%	0	0	9%	0.016	98%	0	0	11%	0.018	97%
MI-LV	0	0	2%	0.010	94%	0	0	2%	0.010	94%	0	0	3%	0.011	94%
BSEM	0	3	8%	0.025	95%	0	1	10%	0.032	95%	0	2	12%	0.037	96%
N = 300, Mprop = 30%															
RFIML	0	0	1%	0.012	95%	0	0	2%	0.014	95%	0	0	2%	0.015	95%
MI-MVN	0	0	1%	0.013	95%	2	2	3%	0.015	95%	0	0	3%	0.016	94%
cat-DWLS	0	0	3%	0.012	96%	0	0	5%	0.018	97%	0	0	6%	0.020	97%
MICE-LOGIT	1	1	3%	0.012	95%	12	12	6%	0.018	94%	24	26	7%	0.025	92%
MICE-RF	0	0	5%	0.013	98%	4	7	19%	0.031	99%	11	15	23%	0.042	99%
MI-LV	0	0	3%	0.012	95%	0	0	3%	0.013	95%	0	0	4%	0.014	95%
BSEM	0	2	14%	0.039	94%	0	0	23%	0.072	94%	0	1	21%	0.061	95%
N = 600, Mprop = 15%															
RFIML	0	0	1%	0.005	95%	0	0	1%	0.006	96%	0	0	1%	0.006	95%
MI-MVN	0	0	1%	0.005	94%	0	0	1%	0.006	95%	0	0	1%	0.006	94%
cat-DWLS	0	0	1%	0.005	95%	0	0	-1%	0.005	96%	0	0	-1%	0.006	96%
MICE-LOGIT	0	0	1%	0.005	94%	0	0	3%	0.008	91%	0	0	4%	0.011	87%
MICE-RF	0	0	2%	0.005	97%	0	0	8%	0.008	97%	0	0	9%	0.009	97%
MI-LV	0	0	1%	0.005	94%	0	0	2%	0.005	95%	0	0	2%	0.005	94%
BSEM	0	0	4%	0.008	95%	0	0	5%	0.009	96%	0	0	5%	0.009	95%
N = 600, Mprop = 30%															
RFIML	0	0	1%	0.006	95%	0	0	1%	0.007	96%	0	0	2%	0.007	95%
MI-MVN	0	0	1%	0.006	94%	0	0	2%	0.007	95%	0	0	2%	0.007	95%
cat-DWLS	0	0	2%	0.006	95%	0	0	3%	0.008	97%	0	0	3%	0.009	96%
MICE-LOGIT	0	0	2%	0.006	94%	1	1	3%	0.009	91%	0	0	5%	0.013	89%
MICE-RF	0	0	3%	0.006	98%	0	0	16%	0.016	99%	0	0	19%	0.019	98%
MI-LV	0	0	2%	0.006	95%	0	0	2%	0.006	95%	0	0	2%	0.006	95%
BSEM	0	0	5%	0.010	95%	0	1	6%	0.013	95%	0	2	7%	0.014	95%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%.

Table 15. Results for Missing Five-Category Data with Moderately Asymmetrical Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
							N = 300, Mprop = 15%								
RFIML	0	0	1%	0.011	95%	0	0	2%	0.011	95%	0	0	1%	0.013	95%
MI-MVN	0	0	1%	0.011	95%	0	0	2%	0.012	94%	0	0	1%	0.013	94%
cat-DWLS	0	0	3%	0.010	96%	0	0	2%	0.011	96%	0	0	1%	0.013	96%
MICE-LOGIT	0	0	3%	0.010	95%	1	1	6%	0.012	95%	2	2	-1%	0.014	91%
MICE-RF	0	0	4%	0.011	97%	0	0	8%	0.013	97%	0	0	16%	0.026	98%
MI-LV	0	0	3%	0.010	95%	0	0	4%	0.011	94%	0	0	2%	0.011	95%
BSEM	0	3	8%	0.024	96%	0	1	10%	0.027	95%	0	1	16%	0.048	95%
							N = 300, Mprop = 30%								
RFIML	0	0	1%	0.013	95%	0	0	3%	0.015	94%	0	0	2%	0.017	94%
MI-MVN	0	0	1%	0.014	94%	1	1	3%	0.016	94%	16	16	3%	0.018	95%
cat-DWLS	0	0	3%	0.013	95%	0	0	5%	0.017	96%	0	0	8%	0.025	97%
MICE-LOGIT	1	1	5%	0.013	95%	363	-	-	-	-	2	2	-1%	0.016	91%
MICE-RF	1	1	6%	0.014	98%	0	0	14%	0.024	99%	282	-	-	-	-
MI-LV	0	0	4%	0.013	95%	0	0	5%	0.014	95%	0	0	4%	0.014	95%
BSEM	0	1	14%	0.044	94%	0	1	17%	0.051	93%	0	1	45%	0.168	95%
							N = 600, Mprop = 15%								
RFIML	0	0	1%	0.006	95%	0	0	2%	0.006	95%	0	0	1%	0.006	95%
MI-MVN	0	0	1%	0.006	94%	0	0	2%	0.006	95%	0	0	1%	0.007	94%
cat-DWLS	0	0	2%	0.005	95%	0	0	1%	0.006	95%	0	0	-1%	0.006	97%
MICE-LOGIT	0	0	2%	0.005	95%	0	0	3%	0.006	94%	0	0	-1%	0.007	90%
MICE-RF	0	0	3%	0.005	97%	0	0	7%	0.007	97%	0	0	14%	0.013	97%
MI-LV	0	0	2%	0.005	95%	0	0	3%	0.005	94%	0	0	1%	0.005	94%
BSEM	0	0	5%	0.009	95%	0	0	5%	0.009	95%	0	1	7%	0.011	96%
							N = 600, Mprop = 30%								
RFIML	0	0	1%	0.006	95%	0	0	3%	0.007	95%	0	0	0%	0.008	95%
MI-MVN	0	0	1%	0.007	95%	0	0	3%	0.008	95%	0	0	1%	0.009	95%
cat-DWLS	0	0	3%	0.006	95%	0	0	3%	0.008	96%	0	0	2%	0.010	97%
MICE-LOGIT	0	0	3%	0.006	94%	2	2	8%	0.011	92%	0	0	-1%	0.008	91%
MICE-RF	0	0	4%	0.007	98%	0	0	12%	0.012	98%	10	12	30%	0.038	99%
MI-LV	0	0	3%	0.006	95%	0	0	4%	0.007	95%	0	0	1%	0.007	94%
BSEM	0	1	6%	0.011	95%	0	1	7%	0.013	95%	0	2	13%	0.023	95%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%. Bias, MSE and CIC are not computed if Convergence Failures ≥ 100.

Table 16. Results for Missing Five-Category Data with Severely Asymmetrical Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	0	0	0%	0.013	95%	0	0	1%	0.013	96%	1	2	3%	0.018	95%
MI-MVN	0	0	1%	0.013	93%	0	0	2%	0.013	95%	10	13	3%	0.018	94%
cat-DWLS	0	0	7%	0.013	95%	0	0	9%	0.014	96%	0	0	8%	0.019	97%
MICE-LOGIT	0	0	7%	0.013	95%	0	0	8%	0.014	96%	0	0	-1%	0.015	90%
MICE-RF	0	0	7%	0.014	97%	0	0	4%	0.011	97%	201	-	-	-	-
MI-LV	0	0	7%	0.013	95%	0	0	7%	0.013	96%	0	0	8%	0.016	95%
BSEM	0	1	18%	0.045	95%	0	4	16%	0.038	96%	0	2	42%	0.140	95%
N = 300, Mprop = 30%															
RFIML	0	0	1%	0.016	95%	0	0	2%	0.016	96%	0	0	3%	0.027	94%
MI-MVN	2	2	1%	0.017	94%	0	0	3%	0.017	95%	360	-	-	-	-
cat-DWLS	0	0	7%	0.017	95%	0	0	14%	0.021	97%	1	3	19%	0.044	98%
MICE-LOGIT	0	0	8%	0.017	94%	0	0	13%	0.021	95%	1	1	-5%	0.015	90%
MICE-RF	3	3	9%	0.018	99%	0	0	5%	0.013	98%	990	-	-	-	-
MI-LV	0	0	8%	0.017	95%	0	0	8%	0.016	95%	0	0	9%	0.024	96%
BSEM	0	1	34%	0.116	95%	0	2	25%	0.079	95%	0	0	87%	0.427	94%
N = 600, Mprop = 15%															
RFIML	0	0	1%	0.007	95%	0	0	2%	0.007	95%	0	0	2%	0.008	95%
MI-MVN	0	0	1%	0.007	94%	0	0	2%	0.007	94%	0	0	2%	0.009	94%
cat-DWLS	0	0	6%	0.007	94%	0	0	9%	0.007	94%	0	0	5%	0.009	96%
MICE-LOGIT	0	0	6%	0.007	94%	0	0	6%	0.007	94%	0	0	-3%	0.008	88%
MICE-RF	0	0	6%	0.007	97%	0	0	3%	0.006	96%	6	6	24%	0.028	97%
MI-LV	0	0	6%	0.007	94%	0	0	6%	0.007	94%	0	0	6%	0.008	95%
BSEM	0	1	10%	0.013	95%	0	0	10%	0.011	96%	0	1	15%	0.022	96%
N = 600, Mprop = 30%															
RFIML	0	0	1%	0.008	95%	0	0	3%	0.008	96%	0	0	1%	0.012	94%
MI-MVN	0	0	1%	0.008	94%	0	0	3%	0.008	95%	12	12	1%	0.014	94%
cat-DWLS	0	0	6%	0.008	95%	0	0	13%	0.012	95%	0	0	14%	0.018	98%
MICE-LOGIT	0	0	7%	0.008	94%	0	0	8%	0.009	94%	0	0	-3%	0.009	88%
MICE-RF	0	0	8%	0.009	98%	0	0	3%	0.006	98%	978	-	-	-	-
MI-LV	0	0	7%	0.009	95%	0	0	6%	0.008	95%	0	0	7%	0.011	95%
BSEM	0	1	12%	0.017	96%	0	0	12%	0.016	96%	0	1	36%	0.095	95%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%. Bias, MSE and CIC are not computed if Convergence Failures ≥ 100.

Table 17. Results for Missing Seven-Category Data with Symmetric Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	0	0	1%	0.009	95%	0	0	2%	0.010	95%	0	0	2%	0.011	95%
MI-MVN	0	0	1%	0.010	95%	0	0	2%	0.011	95%	0	0	2%	0.011	94%
cat-DWLS	0	0	2%	0.009	97%	0	0	-2%	0.010	96%	0	0	-1%	0.011	96%
MICE-LOGIT	0	0	2%	0.009	94%	6	7	5%	0.014	94%	22	23	6%	0.019	91%
MICE-RF	0	0	3%	0.010	96%	0	0	10%	0.016	97%	0	0	11%	0.018	97%
MI-LV	0	0	2%	0.009	94%	0	0	2%	0.009	95%	0	0	2%	0.010	94%
BSEM	0	1	10%	0.024	95%	0	0	14%	0.037	94%	0	0	15%	0.035	95%
N = 300, Mprop = 30%															
RFIML	0	0	1%	0.011	95%	0	0	2%	0.013	95%	0	0	2%	0.014	95%
MI-MVN	0	0	1%	0.012	95%	2	2	3%	0.014	94%	0	0	3%	0.015	94%
cat-DWLS	0	0	2%	0.011	97%	0	0	1%	0.017	97%	0	0	3%	0.019	97%
MICE-LOGIT	0	0	3%	0.012	95%	12	13	5%	0.017	95%	63	63	6%	0.023	93%
MICE-RF	0	0	5%	0.012	98%	4	5	19%	0.031	99%	5	8	22%	0.038	99%
MI-LV	0	0	3%	0.011	94%	0	0	3%	0.012	95%	0	0	3%	0.013	93%
BSEM	0	0	6%	0.028	93%	0	0	10%	0.048	91%	0	0	12%	0.050	92%
N = 600, Mprop = 15%															
RFIML	0	0	1%	0.005	95%	0	0	1%	0.005	95%	0	0	1%	0.005	95%
MI-MVN	0	0	1%	0.005	94%	0	0	1%	0.005	95%	0	0	2%	0.005	95%
cat-DWLS	0	0	1%	0.005	95%	0	0	-3%	0.006	96%	0	0	-3%	0.006	96%
MICE-LOGIT	0	0	1%	0.005	94%	0	0	4%	0.007	92%	0	0	5%	0.009	90%
MICE-RF	0	0	2%	0.005	96%	0	0	9%	0.008	96%	0	0	10%	0.009	96%
MI-LV	0	0	1%	0.005	94%	0	0	2%	0.005	94%	0	0	2%	0.005	94%
BSEM	0	0	3%	0.008	95%	0	0	4%	0.009	95%	0	0	4%	0.009	94%
N = 600, Mprop = 30%															
RFIML	0	0	1%	0.006	95%	0	0	2%	0.006	95%	0	0	2%	0.007	95%
MI-MVN	0	0	1%	0.006	94%	0	0	2%	0.007	95%	0	0	2%	0.007	95%
cat-DWLS	0	0	2%	0.006	96%	0	0	-1%	0.008	97%	0	0	-1%	0.008	97%
MICE-LOGIT	0	0	2%	0.006	94%	0	0	4%	0.008	93%	0	0	5%	0.011	90%
MICE-RF	0	0	4%	0.006	98%	0	0	17%	0.015	98%	0	0	19%	0.018	0.984
MI-LV	0	0	2%	0.005	95%	0	0	2%	0.006	95%	0	0	2%	0.006	95%
BSEM	0	1	3%	0.009	95%	0	0	5%	0.011	95%	0	0	5%	0.012	94%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%.

Table 18. Results for Missing Seven-Category Data with Moderately Asymmetrical Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
N = 300, Mprop = 15%															
RFIML	0	0	1%	0.010	95%	0	0	2%	0.011	95%	0	0	0%	0.012	95%
MI-MVN	0	0	1%	0.010	95%	0	0	2%	0.011	94%	0	0	1%	0.012	94%
cat-DWLS	0	0	3%	0.009	97%	0	0	0%	0.011	97%	0	0	-1%	0.012	97%
MICE-LOGIT	0	0	3%	0.010	94%	215	-	-	-	-	13	14	2%	0.015	91%
MICE-RF	0	0	4%	0.010	96%	0	0	8%	0.013	97%	0	0	15%	0.024	97%
MI-LV	0	0	3%	0.009	94%	0	0	4%	0.010	94%	0	0	1%	0.010	93%
BSEM	0	0	11%	0.025	95%	0	0	13%	0.032	95%	0	2	18%	0.043	95%
N = 300, Mprop = 30%															
RFIML	0	0	1%	0.012	95%	0	0	3%	0.014	95%	0	0	0%	0.014	94%
MI-MVN	0	0	1%	0.013	94%	0	0	3%	0.015	94%	3	4	1%	0.015	95%
cat-DWLS	0	0	3%	0.012	97%	0	0	1%	0.016	98%	0	0	4%	0.020	97%
MICE-LOGIT	0	0	4%	0.012	95%	466	-	-	-	-	49	52	1%	0.017	93%
MICE-RF	0	0	5%	0.013	98%	1	1	15%	0.023	99%	60	82	31%	0.058	99%
MI-LV	0	0	3%	0.011	94%	0	0	5%	0.014	95%	0	0	1%	0.012	94%
BSEM	0	2	13%	0.039	93%	0	0	15%	0.048	93%	0	0	26%	0.079	93%
N = 600, Mprop = 15%															
RFIML	0	0	1%	0.005	95%	0	0	2%	0.006	95%	0	0	0%	0.006	94%
MI-MVN	0	0	1%	0.005	94%	0	0	2%	0.006	94%	0	0	0%	0.006	94%
cat-DWLS	0	0	2%	0.005	96%	0	0	-1%	0.005	97%	0	0	-3%	0.006	96%
MICE-LOGIT	0	0	2%	0.005	94%	308	-	-	-	-	1	1	-1%	0.007	90%
MICE-RF	0	0	3%	0.005	97%	0	0	7%	0.007	97%	0	0	13%	0.011	96%
MI-LV	0	0	2%	0.005	94%	0	0	3%	0.005	95%	0	0	0%	0.005	93%
BSEM	0	0	3%	0.008	95%	0	0	4%	0.009	95%	0	0	5%	0.009	95%
N = 600, Mprop = 30%															
RFIML	0	0	1%	0.006	95%	0	0	3%	0.007	96%	0	0	-1%	0.007	95%
MI-MVN	0	0	1%	0.006	94%	0	0	3%	0.007	95%	0	0	0%	0.007	95%
cat-DWLS	0	0	2%	0.006	96%	0	0	-1%	0.008	97%	0	0	-1%	0.009	97%
MICE-LOGIT	0	0	3%	0.006	94%	513	-	-	-	-	2	4	-1%	0.008	90%
MICE-RF	0	0	4%	0.006	98%	0	0	13%	0.012	98%	1	2	27%	0.029	99%
MI-LV	0	0	2%	0.006	95%	0	0	4%	0.007	95%	0	0	0%	0.005	94%
BSEM	0	0	4%	0.009	95%	0	1	5%	0.012	95%	0	1	6%	0.013	95%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%. Bias, MSE and CIC are not computed if Convergence Failures ≥ 100.

Table 19. Results for Missing Seven-Category Data with Severely Asymmetrical Thresholds

	MCAR					MAR-Head					MAR-Tail				
	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC	Conv Failures	Conv Failures + Outliers	Bias	MSE	CIC
							N = 300, Mprop = 30%								
RFIML	0	0	1%	0.011	95%	0	0	1%	0.012	95%	0	0	2%	0.015	95%
MI-MVN	0	0	1%	0.012	94%	0	0	2%	0.012	95%	0	0	2%	0.015	94%
cat-DWLS	0	0	5%	0.011	96%	0	0	7%	0.012	96%	0	0	3%	0.015	97%
MICE-LOGIT	0	0	6%	0.011	94%	0	0	8%	0.013	94%	0	0	-2%	0.013	91%
MICE-RF	0	0	7%	0.012	97%	0	0	4%	0.010	96%	65	71	23%	0.044	98%
MI-LV	0	0	6%	0.011	94%	0	0	5%	0.011	94%	0	0	5%	0.013	94%
BSEM	0	0	17%	0.042	95%	0	0	13%	0.034	95%	0	0	34%	0.099	94%
							N = 300, Mprop = 30%								
RFIML	0	0	1%	0.014	95%	0	0	3%	0.015	95%	0	0	2%	0.021	94%
MI-MVN	0	0	1%	0.015	94%	0	0	3%	0.016	94%	94	102	3%	0.022	94%
cat-DWLS	0	0	5%	0.014	96%	0	0	8%	0.020	97%	0	0	9%	0.029	97%
MICE-LOGIT	0	0	7%	0.015	94%	19	20	17%	0.027	95%	0	0	-5%	0.013	91%
MICE-RF	1	2	10%	0.015	99%	0	0	9%	0.014	98%	981	-	-	-	-
MI-LV	0	0	6%	0.014	94%	0	0	7%	0.015	95%	0	0	7%	0.017	95%
BSEM	0	0	18%	0.050	94%	0	0	16%	0.048	93%	0	1	45%	0.163	92%
							N = 600, Mprop = 15%								
RFIML	0	0	1%	0.006	95%	0	0	2%	0.006	96%	0	0	1%	0.007	95%
MI-MVN	0	0	1%	0.006	94%	0	0	2%	0.006	95%	0	0	1%	0.007	94%
cat-DWLS	0	0	4%	0.006	95%	0	0	6%	0.006	96%	0	0	1%	0.007	97%
MICE-LOGIT	0	0	5%	0.006	94%	0	0	5%	0.006	95%	0	0	-2%	0.007	89%
MICE-RF	0	0	5%	0.006	97%	0	0	3%	0.005	96%	0	0	19%	0.020	97%
MI-LV	0	0	5%	0.006	94%	0	0	4%	0.006	95%	0	0	4%	0.006	95%
BSEM	0	0	9%	0.011	95%	0	0	8%	0.010	96%	0	0	12%	0.014	96%
							N = 600, Mprop = 30%								
RFIML	0	0	1%	0.007	95%	0	0	3%	0.007	96%	0	0	0%	0.009	95%
MI-MVN	0	0	1%	0.007	94%	0	0	3%	0.008	95%	1	1	1%	0.009	95%
cat-DWLS	0	0	5%	0.007	95%	0	0	6%	0.009	97%	0	0	4%	0.012	98%
MICE-LOGIT	0	0	5%	0.007	94%	0	0	9%	0.009	94%	0	0	-3%	0.007	90%
MICE-RF	0	0	7%	0.008	98%	0	0	7%	0.007	98%	875	-	-	-	-
MI-LV	0	0	5%	0.007	94%	0	0	6%	0.007	95%	0	0	4%	0.007	95%
BSEM	0	1	7%	0.012	95%	0	0	7%	0.012	95%	0	4	9%	0.021	94%

Note. Values are highlighted if Convergence Failures ≥ 100, Bias ≥ 15%, MSE ≥ 2 times RFIML's MSE under the same condition, or CIC < 90%. Bias, MSE and CIC are not computed if Convergence Failures ≥ 100.

Chapter 7: Discussion

The purpose of this dissertation is to evaluate the currently available methods to deal with missing non-normal data and to identify the best methods to handle two types of missing non-normal data (i.e., continuous data and ordinal data) across a broad range of conditions, using simulation studies. The results of the simulation studies are discussed with respect to the seven research questions raised in Chapters 5 and 6.

Methods for Missing Continuous Non-Normal Data

Question 1: To what extent are the normal-theory-based MI-MVN and BSEM robust to non-normal continuous data?

MI-MVN was found quite robust to mild and moderate normality; it performed identically to RFIML in most conditions and performed slightly better than RFIML under MAR-Tail. However, for severely non-normal data, the CICs from MI-MVN fell below 90% when sample size was small (N = 300), regardless of the other factors (i.e., missing data mechanism and missing data proportion). These findings are partially consistent with Demirtas, et al. (2008), which examined mild non-normality and claimed that MI-MVN was robust to non-normality. The findings also agree with the conclusion drawn by Yuan et al. (2012), that is, when the underlying distribution of the ordinal data had a heavy tail, the MI-MVN tended to be less reliable than RFIML.

The biases and MSEs yielded by normal-BSEM were comparable to RFIML across different conditions. However, the credible CI coverage from normal-BSEM was too low when data were severely non-normal even when data were completed, or when missing data occurred on the tail of the data distribution with moderate non-normality. Therefore, normal-BSEM seemed to be only robust to mild non-normality or moderate normality if missing data is MCAR.

Question 2: How are the methods influenced by sample size, degree of non-normality, missing data mechanism and missing data proportion?

RFIML was mainly influenced by the degree of non-normality and missing data mechanism. Generally speaking, the rescaling strategy did not work well for CIC if the tail of data distribution was severely heavy and the heavy tail was truncated by missing data. This problem was alleviated by large sample size or small missing data proportion. The impact of non-normality and missing data mechanism was similar for MI-MVN. The only exception is that MI-MVN could not handle small sample size appropriately under MAR-Tail, which resulted in underestimated standard errors and therefore low CI coverages.

In comparison, MICE-PMM was mainly impacted by the degree of non-normality and sample size. This method tended to be less reliable in both point and standard error estimates under severe non-normality with small sample size. MICE-RF, as a donor-based imputation method, produced deflated standard error estimates under MAR-Tail across all other conditions. Normal-BSEM, as discussed above, was most affected by the degree of non-normality and missing data mechanism. Similar to RFIML, the 95% CI coverage of normal-BSEM tended to be lower than 90% when the population distribution was severely non-normal and missing data were MAR-Tail, and this problem was not alleviated even when sample size increased to 600.

Question 3: Which method performs best under a variety of conditions, with respect to sample size, degree of non-normality, missing data mechanism, and missing data proportion?

If the population distribution was known, then all five methods worked well for mild non-normal data, except for MICE-RF. MICE-RF should to be used with caution, because it was problematic with the large missing data proportion (30%). For moderate non-normality, all methods were comparable under MCAR. If missing data was not MCAR, then RFIML, MI-

MVN, or MICE-PMM may be used when the missing data proportion was small (around 15%), and only MICE-PMM performed well with large missing data proportions (30%). MICE-RF or normal-BSEM are not recommended for all conditions examined in the study.

Similarly, for severe non-normality, most of the methods (except for normal-BSEM) could be used with MCAR and large sample size; if sample size was small or MCAR was not tenable, then MICE-PMM was the safest option. Overall, MICE-PMM provided the best performance under all the examined conditions.

Methods for Missing Ordinal Data

Question 4: Are the continuous-data methods RFIML and MI-MVN applicable to ordinal data? Under what situations and to what extent are the two methods robust to discontinuity?

The results show that the continuous-data methods RFIML and MI-MVN in general worked quite well for ordinal data. RFIML was reliable under various conditions examined in the study for missing ordinal data, except that it failed to converge when the data were dichotomous, the threshold was asymmetrical and the missing data occurred on the heavy tail of the distribution (MAR-Tail).

MI-MVN had convergence problems for dichotomous data and three-category ordinal data with asymmetrical thresholds. Compared to RFIML, MI-MVN required a larger sample size or a smaller proportion of missing data in some of the most difficult situations to converge to admissible solutions. However, for data with five or seven categories, the results from MI-MVN and RFIML were almost identical.

Question 5: Do normal-theory-based BSEM and MI-LV perform well under a broader range of conditions than those examined in Asparouhov and Muthén (2010a)?

Asparouhov and Muthén (2010a) found that normal-BSEM and MI-LV were superior to cat-DWLS for symmetric dichotomous data with $N = 1000$. In this study, the performance of these methods was examined under a broader range of conditions. I found that normal-BSEM performed poorly when the sample size or the number of categories was small. Cat-DWLS did not work with dichotomous data under MAR, however, its performance was greatly improved as the number of categories increased. In comparison, MI-LV was a very reliable method, unless the number of categories was two or three and the thresholds were severely asymmetrical.

Question 6: How are the methods influenced by sample size, degree of non-normality, missing data mechanism and missing data proportion?

As discussed above, RFIML and MI-MVN were least impacted by the design factors. However, they might fail to converge under the most difficult conditions (i.e., number of categories was two or three, the thresholds were severely asymmetrical, and missing data were MAR-Tail). Under the other conditions, the performance of the two methods was stable and reliable. MI-LV was mostly affected by number of categories and asymmetry of thresholds. It produced large biases and MSEs when the number of categories was two or three and the thresholds were severely asymmetrical. Normal-BSEM generally required a large sample size, and the impact of sample size on normal-BSEM was moderated by the number of categories. Among the examined conditions, normal-BSEM only performed well for data with five or seven categories and when sample size was 600.

Cat-DWLS were affected by the number of categories, thresholds, and missing data mechanism. For dichotomous data, cat-DWLS only worked well when the thresholds were symmetric and the missing data mechanism was MCAR. As the number of categories increased, the impact of thresholds and missing data mechanism decreased.

Compared to the other methods, MICE-LOGIT was impacted by the factors in a very different way. It generally worked adequately with small number of categories (2 or 3), except when data were severely asymmetrical. When the number of categories was large (5 or 7), MICE-LOGIT showed large numbers of convergence failures under MAR-Head and moderate asymmetrical thresholds.

The performance of MICE-RF was undesirable under the conditions tested. It only worked acceptably with MCAR data and a small missing data proportion.

Question 7: Which of the seven methods performs best under the examined conditions?

Among the seven methods, RFIML performed the best under almost all the conditions, and was least affected by the design factors. The second optimal method would be MI-MVN combined with RML estimator, which could be the best option if MI has to be used to deal with missing data.

Limitations and Future Directions

The findings and conclusions from this dissertation are limited to the scope of the studies. First, I only examined one type of SEM model and focused on the three latent paths. It would be interesting to know whether the examined methods perform differently when applied to other types of SEMs and for different types of parameters. Second, not all the possible conditions were covered in the two studies. For example, if the sample size was less than 300, the best method (i.e., MICE-PMM) identified in Study I might have convergence problems and become suboptimal. On the other hand, normal-BSEM might perform well with small number of categories if the sample size is greater than 600. Third, this dissertation did not cover all methods for missing non-normal data proposed in the past research. For example, because of the computational complexity of MICE-RF, I only examined one of its variations that was found

most reliable in survival analysis. Thus, the current finding does not necessarily indicate that all variations of MICE-RF will not work well for missing non-normal data in the SEM context. Similarly, only the normal-theory-based BSEM (with Mplus default priors) was included. It would be interesting to further explore other strategies in dealing with missing non-normal data in the Bayesian framework. Fourth, the outcome measures used in the present studies focused on the accuracy and precision of the parameter estimates. The chi-square test statistic and other model fit indices were not covered. Rescaling strategies are often used to correct for the impact of non-normality on model fit indices. However, there is little discussion in the literature in terms of how to appropriately pool the rescaled model fit indices across imputed data sets. A future study is warranted to examine whether the existing pooling approaches work for rescaled fit indices and to develop potential solutions that have better performance. Finally, the two studies examined non-normal continuous and ordinal data separately. In practice, the two types of non-normal data very likely coexist in one model. It would also be interesting to investigate scenarios in which the two types of data need to be simultaneously analyzed.

Conclusion

In this dissertation, I have conducted two studies to evaluate the performance of the methods in handling two types of missing non-normal data, (i.e., missing non-normal continuous data and missing ordinal data). The results agree with the previous research and expand their findings to a broader range of conditions. While comparing these methods in terms of convergence failures, relative bias, mean squared error (MSE), and 95% confidence interval (or credible interval) coverage (CIC), the following conclusions can be reached:

For missing non-normal continuous data, MICE-PMM has the most stable performance under the examined conditions among all five methods. RFIML is generally a good estimator,

except when the population distribution is severely non-normal and the missingness occurs primarily on the heavy tail. MI-MVN with RML is more sensitive to severe non-normality than RFIML. Other than that, it could serve as an alternative to RFIML. Finally, MICE-RF and the normal-theory-based BSEM are not good choices for handling missing non-normal continuous data.

For missing ordinal data, first, no method performs well for severely asymmetrical dichotomous data with MAR-Tail missingness. Other than that condition, the continuous-data method RFIML performs the best among all seven methods, followed by MI-MVN. All the cat-DWLS-based methods (i.e., direct cat-DWLS, MICE-LOGIT, MICE-RF and MI-LV) are sensitive to design factors to some degree, and are inferior to RFIML and MI-MVN. Normal-theory-based BSEM is generally not recommended for missing ordinal data for the examined sample sizes.

References

Allison, P. D. (2000). *Missing data*. Thousand Oaks, CA: Sage.

Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review, 78*(1), 40-64.

Asparouhov, T., & Muthén, B. (2010a). Bayesian analysis of latent variable models using Mplus (Technical report). Los Angeles, CA: Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2010b). Bayesian analysis using Mplus: Technical implementation (Technical appendix). Los Angeles, CA: Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2010c). Multiple imputation with Mplus. *Technical Report. www.statmodel.com*.

Asparouhov, T., & Muthén, B. (2010d). Simple second order chi-square correction (Technical Appendix). *Retrieved from http://www.statmodel.com/download/WLSMV_new_chi21.pdf*.

Asparouhov, T., & Muthén, B. (2010e). Weighted Least Squares Estimation with Missing Data (Technical appendix). Los Angeles, CA: Muthén & Muthén.

Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16*(1), 78-117.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*(1), 62-83.

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*(2), 347-357.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing, 6*(2), Bernaards.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16-29.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-38.

Demirtas, H., Freels, S. A., & Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation, 78*(1), 69-84.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*(3), 327-346.

Di Zio, M., & Guarnera, U. (2009). Semiparametric predictive mean matching. *AStA Advances in Statistical Analysis, 93*(2), 175-186.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*(2), 309-326.

Doove, L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational statistics & data analysis, 72*, 92-104.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Enders, C. K. (2001a). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*(1), 128-141.

Enders, C. K. (2001b). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6*(4), 352-370.

Enders, C. K. (2010). *Applied missing data analysis*: The Guilford Press.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods, 16*(1), 1-16.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430-457.

Fan, X., & Wang, L. (1998). Effects of potential confounding factors on fit indices and parameter estimates for true and misspecified SEM models. *Educational and Psychological Measurement, 58*(5), 701-735.

Ferrari, P. A., & Barbiero, A. (2012). Simulating ordinal data. *Multivariate Behavioral Research, 47*(4), 566-589.

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal, 4*(2), 87-107.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269-314). Greenwich, CT: Information Age Publishing.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521-532.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*(4), 625-641.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*(6), 721-741.

Geyer, C. (2011). Introduction to Markov Chain Monte Carlo. In A. G. Steve Brooks, Galin Jones, Xiao-Li Meng (Ed.), *Handbook of Markov Chain Monte Carlo* (pp. 3-48). Boca Raton, FL: Chapman & Hall/CRC.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical Science,* 473-483.

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research, 47*(1), 1-25.

Gilks, W. R. (2005). *Markov chain monte carlo*: Wiley Online Library.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206-213.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 4*(2), 108-120.

Heitjan, D. F., & Little, R. J. (1991). Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society C, 40*(1), 13-29.

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science, 54*(2), 561-581.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software, 45*(7), 1-47.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329-367.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*(2), 351-362.

Iacus, S., & Porro, G. (2006). Missing data imputation, classification, prediction and average treatment effect estimation via Random Recursive Partitioning. *UNIMI-Research Papers in Economics, Business, and Statistics. Statistics and Mathematics*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*: Springer.

Kaplan, D., & Depaoli, S. (2012). Bayesian Structural Equation Modeling. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650-673). New York, NY: Guilford Press.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review, 95*(1), 49-69.

Koller-Meinfelder, F. (2010). *Analysis of Incomplete Survey Data–Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching.* PhD thesis, Otto-Friedrich-University, Bamberg.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. UK: John Wiley & Sons.

Lee, S.-Y., & Song, X.-Y. (2004a). Bayesian model comparison of nonlinear structural equation models with missing continuous and ordinal categorical data. *British Journal of Mathematical and Statistical Psychology, 57*(1), 131-150.

Lee, S.-Y., & Song, X.-Y. (2004b). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*(4), 653-686.

Lee, S.-Y., & Song, X.-Y. (2012). *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*: John Wiley & Sons.

Lee, S.-Y., & Xia, Y.-M. (2008). A robust Bayesian approach for structural equation models with missing data. *Psychometrika, 73*(3), 343-364.

Li, K.-H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association, 86*(416), 1065-1073.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R news, 2*(3), 18-22.

Lindley, D. V., & Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-41.

Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519-530.

Mardia, K. V. (1985). Mardia's test of multinormality. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 5, pp. 217-221). New York: Wiley.

Marshall, A., Altman, D. G., & Holder, R. L. (2010). Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. *BMC Medical Research Methodology, 10*(112). Retrieved from http://www.biomedcentral.com/1471-2288/10/112

Meng, X.-L., & Rubin, D. B. (1992). Performing Likelihood Ratio Tests with Multiply-Imputed Data Sets. *Biometrika, 79*(1), 103-111.

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology, 14*. Retrieved from http://www.biomedcentral.com/1471-2288/14/75

Muthén, B. (2000, March, 28). Re: Underlying normality and polychoric correlations [Online forum comment]. Retrieved from http://www.statmodel.com/cgi-bin/discus/discus.cgi?pg=prev&topic=23&page=73

Muthén, B., du Toit, S. H., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika, 75*, 1-45.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*(2), 171-189.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*(3), 431-462.

Muthén, L. K., & Muthén, B. O. (2008-2012). Mplus user's guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7*(4), 557-595.

Palomo, J., Dunson, D. B., & Bollen, K. (2011). Bayesian structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models*. Amsterdam: Elsevier.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354-373.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresuonse in Surveys*. New York: J. Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 473-489.

SAS Institute Inc. (2010). Introduction to Bayesian Analysis Procedures. *SAS/STAT® 9.22 User's Guide*. Cary, NC: SAS Institute Inc.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. v. Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Savalei, V. (2014). Understanding Robust Corrections in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(1), 149-160.

Savalei, V., & Bentler, P. M. (2005). A statistically justified pairwise ML method for incomplete nonnormal data: A comparison with direct ML and pairwise ADF. *Structural Equation Modeling, 12*(2), 183-214.

Savalei, V., & Falk, C. F. (2014). Robust Two-Stage Approach Outperforms Robust Full Information Maximum Likelihood With Incomplete Nonnormal Data. *Sructural Equation Modeling: A Multidisciplinary Journal, 21*(2), 280-302. doi: 10.1080/10705511.2014.882692

Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from FIML. *Structural Equation Modeling, 19*, 477-494.

Schafer, J. L. (2010). *Analysis of incomplete multivariate data*: CRC press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schenker, N., & Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational statistics & data analysis*, *22*(4), 425-446.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American Journal of Epidemiology, 179*(6), 764-774.

Shapiro, A. (1985). Asymptotic equivalence of minimum discrepancy function estimators to G.L.S. estimators. *South African Statistical Journal*, *19*(1), 73-81.

Song, X.-Y., & Lee, S.-Y. (2002). Analysis of structural equation model with ignorable missing continuous and polytomous data. *Psychometrika, 67*(2), 261-288.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112-118.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323-348.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*(398), 528-540.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48*(3), 465-471.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*(3), 219-242.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049-1064.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1-67.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine, 30*(4), 377-399.

Wu, W., Jia, F., & Enders, C. (2015). A Comparison of Imputation Strategies for Ordinal Missing Data on Likert Scale Variables. *Multivariate behavioral research*, 50(5), 484-503.

Yuan, K.-H., & Bentler, P. M. (2000). Three Likelihood-Based Methods For Mean and Covariance Structure Analysis With Nonnormal Missing Data. *Sociological methodology*, *30*(1), 165-200.

Yuan, K.-H., Lambert, P. L., & Fouladi, R. T. (2004). Mardia's multivariate kurtosis with missing data. *Multivariate Behavioral Research, 39*(3), 413-437.

Yuan, K. H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distribution conditions. *Sociological Methods & Research, 41*(4), 598-629.

Zopluoglu, C. (2013). Generating multivariate non-normal variables [Computer program]. Retrieved from http://sites.education.miami.edu/zopluoglu/software-programs